Symptoms and depression: it's time to break up

Wanders, R.B.K.

Symptoms and depression: it's time to break up

Rob B.K. Wanders

Symptoms and depression: it's time to break up

Cover design by Rene Pronk
Lay-out by Nikki Vermeulen, Ridderprint BV
Printed by Ridderprint BV, www.ridderprint.nl

ISBN: 978-90-367-9582-1 (printed version)
ISBN: 978-90-367-9581-4 (digital version)

rijksuniversiteit
groningen

Symptoms and depression: it's time to break up

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 27 February 2017 at 14.30 hours

by

Robert Bernard Klaas Wanders

born on 21 June 1986
in Veendam

Supervisors
Prof. P. de Jonge
Prof. R.R. Meijer

Co-supervisor
Dr. K.J. Wardenaar

Assessment Committee
Prof. K. Sijtsma
Prof. A.H. Schene
Prof. R.C. Oude Voshaar

TABLE OF CONTENTS

CHAPTER 1

Introduction

DEPRESSION

Depression is one of the leading causes of disability worldwide[1,2], with economic costs that are among the highest of all brain disorders in Europe[3]. Over 1 out of 7 persons develops a mood disorder in their lifetime[4], which can have a chronic or fluctuating course, with recurrence rates up to 85% in specialized care[5]. People with a mood disorder experience disability in social, physical, and occupational domains[6–9]. Furthermore, major depressive disorder is the highest risk factor for suicide attempts and completed suicide[10]. Most people do not receive treatment[11], and for those who do, treatment is often inefficient[12], leaving over 60% of the disease burden intact[13]. The resources required to address these conditions are inadequate, unequally distributed, and inefficiently used[14]. In the US, jails and prisons have now become the largest mental health-care facilities, leading to cries of desperation to bring back the asylum[15]. The problem of depression seems to become bigger, while there is no indication that any viable solutions will be available in the near future.

THIS THESIS

This thesis deploys a data-driven approach to improve the troublesome situation surrounding depression, as it can: (i) increase the understanding of individual and symptom differences[16–18] by empirical evaluation of the depression construct[19,20], (ii) can lead to higher precision in depression assessment and monitoring[21], and (iii) may offer ways to improve effectiveness of mental health care[22,23].

First, no patient or symptom is the same: individuals differ largely in how they respond to treatment[24], in how their depression develops over time[25], and in their presented symptom patterns[26], where each depressive symptom may play a different role[27–29]. These differences are poorly understood in terms of etiology and are a likely reason for scientific stagnation. An empirical evaluation of the depression construct is needed that is not limited to current diagnostic schemes, and that investigates a broader set of symptoms (e.g. anxiety[30]), incorporates different sources of clinically relevant factors (e.g. social functioning[31]) and includes subclinical patients[32]. Second, a more advanced data-analytical approach could increase precision in depression assessments, in part due to increased understanding of differences and similarities across depressed patients. This could offer a more accurate source of phenotypic variation and therefore a more valid target for neurobiological, genetic, and psychosocial research. Third, such a data-driven approach could provide a valuable role by personalizing assessments and providing extra feedback information on top of traditionally reported scores.

Central in this thesis is the use of item response theory (IRT) as a set of data-analytical tools to study the relation between symptoms and the construct of depression. IRT does not refer to a single model or technique but is a collection of statistical approaches that have in common that the goal is to model the relation between item responses and the underlying dimension that is assumed to be measured. Within depression research we are designated to the use of questionnaires to assess depression severity, with items that represent depressive symptoms and an underlying dimension assumed to reflect depression severity. IRT therefore provides the ideal set of tools that can (a) yield valuable insights into the heterogeneity of depression and depression as a construct (b) be used to assess psychometric properties of depression questionnaires and achieve increased measurement precision, and (c) can personalize assessments and offer extra individual feedback in clinical care. Since many chapters in this thesis can be seen from each of these three different perspectives, these are each discussed in more detail below.

HETEROGENEITY PERSPECTIVE ON THIS THESIS

"A principle difficulty in the study of the depressive disorders is that the depressive diseases are associated with a great deal of heterogeneity" – Blumenthal, 1971 [33]

Tremendous advances have been made in biological, neurological, and genetic research, which has benefited many medical fields and continues to provide a beacon of hope for the future of mental health research. At the same time, it has been a disappointment that scientific breakthroughs with clinical impact for depression have been largely absent. The heterogeneity of depression is often seen as a key obstacle in research, hampering progress and responsible for the lack of advances in clinical treatment. The simple reason here being that if patients are so unlike each other, why should they have shared causal mechanisms and benefit from the same types of treatment?

The notion that heterogeneity might obscure scientific results is not new, and was already a topic of research several decades ago[33,34]. Studies investigating primary depression versus depression secondary to other disorders already concluded that heterogeneity in the composition of the sample could obscure important associations, and that investigation of depressive symptom scores alone would not suffice[34]. Nowadays, such a conclusion seems to be more relevant and vibrant than ever before. Serious concerns have been raised about the current diagnostic system[35,36], with boundaries of current classifications being non-existent in empirical data[37], and the implied dichotomy between normal and abnormal disordered states being unlikely to reflect the true nature of mental health[38]. To overcome these problems, researchers have called for more empirically identified homogenous subtypes to advance biological, neurological, and genetic research[39,40], and that cut across current classifications[41].

In this thesis the empirical approach of depression is a central idea underlying the addressed research questions that are focused on the investigation of different sources of heterogeneity and aimed to enable better description of the heterogeneity of depression. Analyzing the relation between symptoms and depression in closer detail could be key to advance our understanding of both depression and its heterogeneity. The IRT framework provides a range of different means to investigate depression heterogeneity, disentangling individual depressive symptoms from the overarching disorder.

Depressive symptoms across clinical groups

Perhaps the simplest approach we can take to investigate individual differences is to look at how different observable groups experience different symptoms. Do depressed women experience different symptoms than depressed men? Is depression expressed differently by older people than by younger people? One problem with these questions is that if these groups are directly compared, it is not possible to know whether women really experience different symptoms or that they are just more depressed in general. We would therefore like to account for such differences in severity when comparing symptom patterns. Such an approach could help to answer the more interesting question: do women that are equally depressed as men have different probabilities of experiencing certain symptoms? This type of question can be answered by studying differential item functioning (DIF[42]), where patients from different groups that are equal in their level of depression severity do not have the same probability of experiencing a particular symptom. A well-known example of a symptom that functions differently between groups is 'crying'. Women are more likely to report that they cry more than usual, than equally depressed men[43]. As such, the symptom is less indicative for depression in women than it is in men. This is of course rather obvious, but DIF analyses might provide more insightful results when applied to groupings of greater clinical interest, where differential roles of symptoms are not immediately clear.

Depressive symptoms across latent groups

Instead of above mentioned observable groups, a data-driven approach can be used to study differences and similarities in symptom patterns across patient groups that might be difficult to observe directly in terms of clinical indicators. An illustrative example, and one of the first statistical applications of finding latent subgroups, was on the classification of the sex of halibut flatfish[44]. The issue with halibut is that it is not possible to directly establish whether it is female or male without dissecting the fish. However, male and female halibut differ in terms of length across age, which can be approximated by a mixture of two normal distributions. This allowed for non-intrusive classification and the estimation of the proportion of male and female halibut in commercial catch. Although depressed patients are no fish, such statistical approaches may increase our understanding of differences and similarities in symptom patterns between depressed patients.

Data-driven methods have been used to identify more homogenous classes[45] or symptom dimensions[46] in depression that could aid in our understanding of the specific underlying etiological processes[47] and benefit clinical decision making[48]. These studies have shown promising results[49], but above all, reviews on the topic show a wide diversity in the found factors and classes[45,47]. Despite the empirical approach in these studies, many are still restricted to symptom scores of a single disorder as defined by current diagnostic classifications. Models that investigate a broader set of symptoms and incorporate additional relevant sources of clinically relevant variability may shed more light on this. These models and their variations, whether they are categorical (e.g. LCA[45]), dimensional (e.g. factor analysis[50]) or a combination of both (e.g. factor mixture models[51]), are possible ways to better understand the data[52], and describe and explain depression heterogeneity.

As an alternative to finding different more homogenous subgroups, we can also approach the problem by trying to identify those that make the sample heterogeneous in terms of symptom patterns. To put it in other words, we could identify a group of patients that report symptoms in patterns that are not consistent with how the majority of patients typically report them. An interesting example of a study taking this approach is one that investigated atypical suicide risk by means of so-called person-fit statistics[53]. In most cases, suicidal ideation is observed when many other internalizing symptoms are present (e.g. sad mood, loss of interest), but in a minority of patients suicidality occurs out of the blue. The authors assessed a range of internalizing symptoms in patients that presented with substance-related problems. Patients identified with atypical response patterns according to empirically based person-fit scores, reported suicidal ideation but few other internalizing symptoms. After identification of such an atypical group, it can be further studied to gain new insights into determinants of the atypical profile. The use of person-fit as a tool to investigate the heterogeneity of depression allows for a novel and interesting approach to decompose the sample into a group of patients with typical symptom patterns and a group with symptom patterns that are atypical for depression.

MEASUREMENT PERSPECTIVE ON THIS THESIS

"Measurement scales are not God-given but rather are a matter of convention" – Nunnally, 1967 [54]

The study of heterogeneity of depression is necessarily intertwined with the validity of measuring the construct of depression. If we are to understand individual differences, and find new associations with neurobiological or genetic variables, we need a strong and valid measure of depression[21,55]. The IRT approach of this thesis can yield valuable new information about the validity of depression assessment.

The goal of IRT is to provide a measurement model in which the characteristics of the test are separated from the assessed individuals. Within such a model, the probability of a patient to report a symptom depends on (a) the patient's location on the underlying dimension (e.g. depression severity) and (b) the characteristics of the item (symptom). For example, items may differ in symptom severity (e.g. sad mood vs. suicidal ideation), or their ability to discriminate between patients at different levels of depression (e.g. sad mood will be more informative about depression severity than weight change). In this context, IRT can be used to estimate these item characteristics, and to scale patients accordingly based on their response patterns (experienced symptoms). The models and associated methods applied in this thesis can be psychometrically informative about (a) the quality and structure of depression measures and (b) the validity of depression assessments.

First, an important source of psychometric information comes from the fact that there are many assumptions underlying IRT models that have to be checked, and which provide insights on quality of data and the structure of depression questionnaires[56]. In short, these analyses could inform us about the extent to which: (i) items measure a single depression construct (unidimensionality), (ii) associations between symptoms are solely due to their relation with depression severity (local independence), and (iii) items and associated categories all have an increased likelihood of being reported at higher levels of depression.

Second, self-report depression questionnaires may not validly assess depression severity for all individuals, leading to biased results and over- or underestimation of depression severity. As discussed before, DIF occurs when patients have a different probability of reporting a symptom due to their group membership. This can be informative about the heterogeneity of depression, but may also expose a measurement problem. For example, cardiovascular patients may endorse certain items (e.g. sleep disturbances) on a depression questionnaire as a direct result of their cardiovascular disease or treatment[57]. This could lead to item bias with cardiovascular patients having higher probabilities of endorsing symptoms when adjusted for depression severity. In addition, some patients may show a symptom pattern that is unexpected given the measurement model that holds for the majority of patients (e.g. reporting of severe symptoms without milder symptoms). These can be identified by means of person-fit statistics and indicate a threat to validity, where the total score is not a good reflection of depression severity.

CLINICAL PERSPECTIVE ON THIS THESIS

"To ignore a source of data because it may be misleading would be like ignoring the footprints in the garden because they might not belong to the burglar" – Funder, 2007 [58]

It is not unthinkable that in the near future, we will establish for each individual separately what should be the target for treatment or prevention, all on a level of detail and precision that is higher than is currently possible. Such an advance will not come from a single genetic risk profile, or a test on biological risk factors. Instead, it is more likely that a complex dynamic interplay of all kinds of different factors play a role in each patient. This includes genetic and biological factors, but extents to psychological, behavioral, social and cultural factors[22,23], and that will require many kinds of data to achieve precision. To reach this goal in depression, an emphasis is needed on both the individual needs[59] as well as the supporting role of technology to obtain accurate assessment, detailed monitoring, and individualized feedback[60] in order to tailor interventions and support accordingly[23].

The tools of IRT could play a valuable role in personalizing assessments and providing extra feedback information on top of traditionally used sum scores.

Assessments can be tailored to individual characteristics by utilizing the information that IRT methods summarize about the relation between symptoms and depression severity across different groups. This can be used both to select the right type of symptoms to assess[61] and to adjust for differential roles of symptoms across individuals based on demographic and clinical characteristics[62]. That is, given our earlier example of DIF of the symptom 'crying' between men and women, such a system could weigh the symptom as more severe when a male patient reports to 'cry more than usual' compared to a female patient (or decide not to select the symptom for assessment at all in male respondents).

Clinicians could be provided with more detailed feedback about the reported symptom pattern of a patient on top of the sum score. Person-fit statistics can be used to identify those individuals that report a symptom pattern that is inconsistent with how the majority of patients typically report depression, and that could signal a different need of care. These could both be individuals that report symptoms due to other psychiatric or somatic disorders, as well as individuals that have responded in an unmotivated or careless fashion. In both cases, a feedback report could be generated for the clinician alerting that the measurement does not reflect a typical pattern of depressive symptoms. Such information may prove to be valuable for clinicians, who can than act accordingly. Until now, there have been no studies that evaluated whether providing clinicians with such feedback is informative in clinical practice.

OUTLINE OF THESIS

The empirical approach of this thesis starts in **chapter 2**, where differences in the relation between symptoms and depression are studied by decomposing the sample into those with typical and those with atypical symptom patterns. It was evaluated whether this could indeed be a useful data-analytical tool to reveal sources of depression heterogeneity, and whether identification of atypical patterns could have potential clinical uses in depression assessment. In **chapter 3** this technique was used to investigate symptom reporting in heart patients, and to assess the extent to which these patients report depressive symptoms that do not reflect depression severity.

In **chapter 4** it was investigated how symptom patterns may differ across equally depressed patients with different clinical characteristics, and what the impact of such differences could be on interpretability and comparability of depression scores by adjusting for differential item functioning.

Fundamental issues were addressed in **chapter 5** on the validity of measuring individual depressive symptoms by investigating item and category functioning in a depression questionnaire, and in **chapter 6** on the problem of strong dependencies between symptoms for finding homogenous depression subtypes.

Chapters 7 and 8 describe results from a data-driven method to identify and validate cross-diagnostic subtypes by simultaneously considering symptoms of depression and anxiety, and disability measures, in two large cohorts (Lifelines, n=73,403; NEMESIS, n=5,583).

The empirical approach in this thesis ends with **chapter 9**, where the added value of data-analytical tools is assessed in a real clinical setting, by performing a pilot study on the use of automated feedback based on person-fit statistics in clinical depression care.

REFERENCES

1. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* **386,** 743–800 (2015).
2. Vigo, D., Thornicroft, G. & Atun, R. Estimating the true global burden of mental illness. *Lancet Psychiatry* **3,** 171–178 (2016).
3. Olesen, J. *et al.* The economic cost of brain disorders in Europe. *Eur. J. Neurol.* **19,** 155–162 (2012).
4. Bromet, E. *et al.* Cross-national epidemiology of DSM-IV major depressive episode. *BMC Med.* **9,** 90 (2011).
5. Hardeveld, F., Spijker, J., De Graaf, R., Nolen, W. A. & Beekman, A. T. F. Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatr. Scand.* **122,** 184–191 (2010).
6. Bijl, R. V. & Ravelli, A. Psychiatric morbidity, service use, and need for care in the general population: results of The Netherlands Mental Health Survey and Incidence Study. *Am. J. Public Health* **90,** 602–607 (2000).
7. Kessler, R. C. *et al.* Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **62,** 593–602 (2005).
8. Patten, S. *et al.* Prospective evaluation of the effect of major depression on working status in a population sample. *Can. J. Psychiatry Rev. Can. Psychiatr.* **54,** 841–845 (2009).
9. Wittchen, H. U. *et al.* The size and burden of mental disorders and other disorders of the brain in Europe 2010. *Eur. Neuropsychopharmacol.* **21,** 655–679 (2011).
10. Holma, K. M. *et al.* Incidence and Predictors of Suicide Attempts in DSM–IV Major Depressive Disorder: A Five-Year Prospective Study. *Am. J. Psychiatry* **167,** 801–808 (2010).
11. Muñoz, R. F., Cuijpers, P., Smit, F., Barrera, A. Z. & Leykin, Y. Prevention of Major Depression. *Annu. Rev. Clin. Psychol.* **6,** 181–212 (2010).
12. Kirsch, I. *et al.* Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration. *PLOS Med.* **5,** e45 (2008).
13. Andrews, G., Issakidis, C., Sanderson, K., Corry, J. & Lapsley, H. Utilising survey data to inform public policy: comparison of the cost-effectiveness of treatment of ten mental disorders. *Br. J. Psychiatry* **184,** 526–533 (2004).
14. Saxena, S., Thornicroft, G., Knapp, M. & Whiteford, H. Resources for mental health: scarcity, inequity, and inefficiency. *The Lancet* **370,** 878–889 (2007).
15. Sisti, D. A., Segal, A. G. & Emanuel, E. J. Improving Long-term Psychiatric Care: Bring Back the Asylum. *JAMA* **313,** 243–244 (2015).
16. Goldberg, D. The heterogeneity of 'major depression'. *World Psychiatry* **10,** 226–228 (2011).
17. Wardenaar, K. J. & de Jonge, P. Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC Med.* **11,** 201 (2013).
18. Fried, E. I. & Nesse, R. M. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med.* **13,** 72 (2015).
19. Zimmerman, M., McGlinchey, J. B., Young, D. & Chelminski, I. Diagnosing major depressive disorder I: A psychometric evaluation of the DSM-IV symptom criteria. *J. Nerv. Ment. Dis.* **194,** 158–163 (2006).
20. Lux, V. & Kendler, K. Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria., Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychol. Med. Psychol. Med.* **40, 40,** 1679, 1679–1690 (2010).
21. Reise, S. & Rodriguez, A. Item response theory and the measurement of psychiatric constructs: some empirical and conceptual issues and challenges. *Psychol. Med.* **46,** 2025–2039 (2016).
22. Insel, T. R. The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *Am. J. Psychiatry* **171,** 395–397 (2014).
23. Bickman, L., Lyon, A. R. & Wolpert, M. Achieving Precision Mental Health through Effective Assessment, Monitoring, and Feedback Processes. *Adm. Policy Ment. Health* **43,** 271–276 (2016).
24. Simon, G. E. & Perlis, R. H. Personalized Medicine for Depression: Can We Match Patients With Treatments? *Am. J. Psychiatry* **167,** 1445–1455 (2010).

25. Eaton, W. W. *et al.* Population-Based Study of First Onset and Chronicity in Major Depressive Disorder. *Arch. Gen. Psychiatry* **65,** 513–520 (2008).
26. Fried, E. I. & Nesse, R. M. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *J. Affect. Disord.* **172,** 96–102 (2015).
27. Hoen, P. *et al.* Differential associations between specific depressive symptoms and cardiovascular prognosis in patients with stable coronary heart disease., Differential associations between specific depressive symptoms and cardiovascular prognosis in patients with stable coronary heart disease. *J. Am. Coll. Cardiol. J. Am. Coll. Cardiol.* **56, 56,** 838, 838–844 (2010).
28. Fried, E. I. & Nesse, R. M. The Impact of Individual Depressive Symptoms on Impairment of Psychosocial Functioning. *PLOS ONE* **9,** e90311 (2014).
29. Boschloo, L. *et al.* The Network Structure of Symptoms of the Diagnostic and Statistical Manual of Mental Disorders. *PLOS ONE* **10,** e0137621 (2015).
30. Simms, L. J., Prisciandaro, J. J., Krueger, R. F. & Goldberg, D. P. The structure of depression, anxiety and somatic symptoms in primary care. *Psychol. Med.* **42,** 15–28 (2012).
31. McKnight, P. E. & Kashdan, T. B. Purpose in life as a system that creates and sustains health and well-being: An integrative, testable theory. *Rev. Gen. Psychol.* **13,** 242–251 (2009).
32. Das-Munshi, J. *et al.* Public health significance of mixed anxiety and depression: beyond current classification. *Br. J. Psychiatry* **192,** 171–177 (2008).
33. Blumenthal, M. D. Heterogeneity and Research on Depressive Disorders. *Arch. Gen. Psychiatry* **24,** 524–531 (1971).
34. Weissman, M. M. *et al.* Symptom Patterns in Primary and Secondary Depression: A Comparison of Primary Depressives With Depressed Opiate Addicts, Alcoholics, and Schizophrenics. *Arch. Gen. Psychiatry* **34,** 854–862 (1977).
35. Kendler, K. S. & First, M. B. Alternative futures for the DSM revision process: iteration v. paradigm shift. *Br. J. Psychiatry* **197,** 263–265 (2010).
36. Whooley, O. Nosological reflections: the failure of DSM-5, the emergence of RDoC, and the decontextualization of mental distress. *Soc. Ment. Health* **4,** 92–110 (2014).
37. Krueger, R. F. & Markon, K. E. Reinterpreting Comorbidity: A Model-Based Approach to Understanding and Classifying Psychopathology. *Annu. Rev. Clin. Psychol.* **2,** 111–133 (2006).
38. Kendler, K. S. The dappled nature of causes of psychiatric illness: replacing the organic–functional/hardware–software dichotomy with empirically based pluralism. *Mol. Psychiatry* **17,** 377–388 (2012).
39. Kapur, S., Phillips, A. G. & Insel, T. R. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* **17,** 1174–1179 (2012).
40. Cuthbert, B. N. & Insel, T. R. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* **11,** 126 (2013).
41. Cuthbert, B. N. The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry* **13,** 28–35 (2014).
42. Holland, P. W. & Wainer, H. *Differential Item Functioning*. (Routledge, 2012).
43. Kim, Y., Pilkonis, P. A., Frank, E., Thase, M. E. & Reynolds, C. F. Differential functioning of the Beck Depression inventory in late-life patients: Use of item response theory. *Psychol. Aging* **17,** 379–391 (2002).
44. Hosmer, D. W. A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of a Mixture of Two Normal Distributions Under Three Different Types of Sample. *Biometrics* **29,** 761–770 (1973).
45. van Loo, H. M., de Jonge, P., Romeijn, J.-W., Kessler, R. C. & Schoevers, R. A. Data-driven subtypes of major depressive disorder: a systematic review. *BMC Med.* **10,** 156 (2012).
46. Wardenaar, K. J. Syndromes versus symptoms : towards validation of a dimensional approach of depression and anxiety. (2012).
47. Baumeister, H. & Parker, G. Meta-review of depressive subtyping models. *J. Affect. Disord.* **139,** 126–140 (2012).
48. Andrews, G., Anderson, T. m., Slade, T. & Sunderland, M. Classification of Anxiety and Depressive disorders: problems and solutions. *Depress. Anxiety* **25,** 274–281 (2008).
49. Lamers, F. *et al.* Identifying Depressive Subtypes in a Large Cohort Study: Results From the Netherlands Study of Depression and Anxiety (NESDA). *J. Clin. Psychiatry* **71,** 1582–1589 (2010).

50. Shafer, A. B. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *J. Clin. Psychol.* **62,** 123–146 (2006).
51. Lubke, G. H. & Muthén, B. Investigating Population Heterogeneity With Factor Mixture Models. *Psychol. Methods* **10,** 21–39 (2005).
52. Loken, E. & Molenaar, P. in *Advances in latent variable mixture models* 227–297 (2008).
53. Conrad, K. J. *et al.* Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug Alcohol Depend.* **106,** 92–100 (2010).
54. Nunnaly, J., Berstein, I. & Berge, J. ten. *Psychometric theory*. (McGraw-Hill, 1967).
55. Huys, Q. J. M., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19,** 404–413 (2016).
56. Meijer, R., Tendeiro, J. & Wanders, R. in *Handbook of item response theory modeling: Applications to typical performance assessment* 85–110 (Routledge, 2014).
57. Delisle, V. C., Beck, A. T., Ziegelstein, R. C. & Thombs, B. D. Symptoms of heart disease or its treatment may increase Beck Depression Inventory Scores in hospitalized post-myocardial infarction patients. *J. Psychosom. Res.* **73,** 157–162 (2012).
58. Funder, D. *The personality puzzle*. (W.W. Norton, 2007).
59. Carney, P. H. Information Technology and Precision Medicine. *Semin. Oncol. Nurs.* **30,** 124–129 (2014).
60. Sacchi, L., Lanzola, G., Viani, N. & Quaglini, S. Personalization and Patient Involvement in Decision Support Systems: Current Trends. *Yearb. Med. Inform.* **10,** 106–118 (2015).
61. Gibbons, R. D. *et al.* Development of a Computerized Adaptive Test for Depression. *Arch. Gen. Psychiatry* **69,** 1104–1112 (2012).
62. Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S. & Crane, P. K. Modifying Measures Based on Differential Item Functioning (DIF) Impact Analyses. *J. Aging Health* **24,** 1044–1076 (2012).

CHAPTER 2

Data-driven Atypical Profiles of Depressive Symptoms: Identification and Validation in a Large Cohort

Rob B. K. Wanders, Klaas J. Wardenaar, Brenda W.J.H. Penninx, Rob R. Meijer, Peter de Jonge

Journal of affective disorders 2015, *180:36-43*.

ABSTRACT

Background. Atypical response behavior on depression questionnaires may invalidate depression severity measurements. This study aimed to identify and investigate atypical profiles of depressive symptoms using a data-driven approach based on item response theory (IRT).

Methods. A large cohort of participants completed the Inventory of Depressive Symptomatology self-report (IDS-SR) at baseline (n=2292) and two-year follow-up (n=1971). Person-fit statistics were used to quantify how strongly each patient's observed symptom profile deviated from the expected profile given the group-based IRT model. Identified atypical profiles were investigated in terms of reported symptoms, external correlates and temporal consistency.

Results. Compared to others, atypical responders (6.8%) showed different symptom profiles, with higher 'mood reactivity' and 'suicidal ideation' and lower levels of mild symptoms like 'sad mood'. Atypical responding was associated with more medication use (especially tricyclic antidepressants: OR=1.5), less somatization (OR=0.8), anxiety severity (OR=0.8) and anxiety diagnoses (OR=0.8-0.9), and was shown relatively stable (29.0%) over time.

Limitations. This is a methodological proof-of-principal based on the IDS-SR in outpatients. Implementation studies are needed.

Conclusion. Person-fit statistics can be used to identify patients who report atypical patterns of depressive symptoms. In research and clinical practice, the extra diagnostic information provided by person-fit statistics could help determine if respondents' depression severity scores are interpretable or should be augmented with additional information.

INTRODUCTION

Major Depressive Disorder (MDD) is a burdensome disorder with heterogeneous symptomatology[1–3] and course trajectories[4,5]. This heterogeneity is a likely reason for the persistent lack of comprehensive etiological models for depression[6]. In order to improve this situation, researchers have attempted to identify more homogenous clinical entities (e.g. subtypes) that better capture the variability among depression patients in terms of phenomenology and etiology.

Depression subtypes are based on clinical consensus (e.g. melancholic or atypical depression[7]) or on empirically-derived common patterns of depressive symptoms. The latter have been investigated with latent class analyses (LCA), which has provided interesting insights into the heterogeneity among depressed patients[8,9]. However, a consistent and well-replicated subtyping model to capture all their inter-individual differences has not yet been established[10]. This could be due to the limitations of LCA[11], and sample/design inconsistencies across studies[10]. However, a more basal issue is that the models are based on subjectively reported symptoms that are all assumed to reflect the construct of depression, which is not necessarily true.

Depressive symptoms can be reported for other reasons than the presence of MDD, such as comorbid somatic or psychiatric disorders, the presence of isolated symptoms, secondary gains by over- or underreporting of symptoms and the existence of specific subtypes of depression. This can result in *atypical* profiles of reported depressive symptoms, which means that patients with such patterns do not conform to definitions of depression. For instance, some depressive patients with somatic illness tend to more often endorse somatic-depressive symptoms, leading to patterns of reported symptoms that do not exclusively reflect depression severity[12]. This is problematic for the assessment of depression severity because scores of persons with atypical profiles do not adequately reflect the assumed underlying construct of depression and cannot be classified or scaled accordingly on a depression severity dimension.

The above described heterogeneity of response behavior can be investigated with a data-driven approach based on *person-fit* statistics and item response theory (IRT[13]). Through person-fit statistics, researchers can investigate the extent to which a respondent's observed score pattern deviates from the expected pattern based on a group-based IRT model[14]. A particular pattern of depressive symptoms can be empirically classified as atypical when too many unexpected scores are observed (e.g. reporting severe symptoms but no mild symptoms). This approach allows for a data-driven identification of atypical response profiles, making no a priori assumptions about what these profiles look like. As a result, the technique is not limited to pre-specified depression classifications or subtypes and could yield new insights into variations in depressive symptom reporting.

To our knowledge, only three previous studies have used person-fit analyses in mental health-related research. First, Conijn et al.[15] identified atypical response patterns on health-related outcome measures among clinical outpatients. These patterns were associated with severe psychological distress and psychopathology, including somatoform disorders, psychotic disorders, and substance-related disorders. Second, Woods et al.[16] found that atypical responding on personality questionnaires was associated with personality pathology. These two studies suggest that person-fit statistics can identify atypical response patterns that are reflective of relevant inter-individual differences and do not arise merely due to chance or non-systematic influences (e.g. test behavior). In a third study, Conrad et al.[17] used person-fit analyses to screen for 'atypical suicide risk', using a questionnaire of internalizing symptoms that was administered to patients with substance-related problems. Those that reported suicidality, but no or few other internalizing symptoms were identified as atypical responders. These patients reported suicidality out of the blue, not in the context of severe internalizing symptomatology. By identifying the latter group, this study showed the extra diagnostic information that person-fit statistics could provide on top of traditional compound scores.

This study aimed to use person-fit analyses to investigate symptom reporting on the Inventory of Depressive Symptomatology Self Report (IDS-SR) in a large cohort study. First, person-fit statistics were used to identify persons with atypical response patterns, given the underlying IRT model of depression severity. Second, item-responses in the atypical responders were investigated. Third, associations of atypical response patterns with external factors were investigated. Finally, the consistency of atypical response behavior over time was investigated.

METHODS

PARTICIPANTS AND PROCEDURES

Data came from the Netherlands Study of Depression and Anxiety (NESDA), a large scale longitudinal cohort study among 2981 adult participants (aged 18-65; 1002 men, 1979 women). Participants were recruited in the general population (19%), primary care (54%), and secondary care (27%). Exclusion criteria were not being fluent in Dutch and/or having a primary diagnosis of bipolar disorder, obsessive compulsive disorder, psychotic disorder, or severe addiction disorder. A follow-up assessment was conducted after two years with a response rate of 87.1% (n=2596). Details about the rationale, objectives, and methods of the study can be found in Penninx et al.[18].

All participants had a face-to-face assessment session with a trained research assistant, consisting of a standardized psychiatric and demographic interview, biomedical measurements, a blood-draw and a battery of self-report questionnaires. The protocol of the NESDA study was approved by the Ethical Committees of all participating universities. All participants signed informed consent.

Data for the current study came from the baseline assessment and the two-year follow up. Only participants with a lifetime anxiety or depression diagnosis (n=2329; 78.1%) were included. Of these, 1971 (84.6%) provided follow up data. From these samples, patients with >5 missing values on the IDS-SR were excluded, leading to a baseline sample of 2292 patients and a follow-up sample of 1942 patients.

MEASURES

Depressive symptoms

The IDS-SR[19] (Rush et al., 1996) is a self-report questionnaire consisting of 30 items rated on a 4-point (0-3) Likert scale. A participant could either endorse 'appetite increase' or 'appetite decrease' and either 'weight increase' or 'weight decrease'. Therefore, these items were combined respectively into compound 'appetite change' and 'weight change' items. The IDS-SR assesses all DSM-IV criterion symptoms for MDD and the most commonly associated symptoms (e.g. anxiety, irritability).

External variables

As no previous studies investigated person-fit in depression, there were no a priori hypotheses about factors that might be associated with atypical depressive symptom reporting. Therefore, a data-mining strategy was used to investigate which out of a wide range of explanatory variables predicted atypical symptom reporting. The used external variables included socio-demographic, clinical, and biological factors. Socio-demographic factors (gender, age, healthcare setting, years of education and north-European ancestry) were assessed at baseline. The Composite International Diagnostic Interview (CIDI, WHO version 2.1) was conducted at baseline to assess the presence of lifetime and current (past six months) DSM-IV diagnoses of MDD, dysthymia, social phobia, generalized anxiety disorder, panic disorder and agoraphobia, alcohol use disorder (alcohol abuse/alcohol dependence). Dichotomous DSM-IV MDD subtype specifiers (atypical and melancholic) were derived from the IDS-SR and calculated regardless of CIDI diagnosis. Anxiety severity was measured with the 21-item Beck Anxiety Inventory (BAI[20]). Both the continuous total BAI score and a categorical BAI severity indicator (≥10:mild, ≥19:moderate, ≥30:severe[20]) were investigated. Manic symptoms were assessed using the 15-item Mood Disorder Questionnaire (MDQ[21]). Both the continuous total scale score

and dichotomous indicator of positive screening (MDQ≥7) for (hypo)manic episode were used. Of the Four Dimensional Symptom Questionnaire (4DSQ[22]), the distress (16 items, range 0-32) and somatization (16 items, range 0-31) scales were used as continuous indicators. In addition, a dichotomous somatization indicator (somatization≥11) was used. Big-Five personality traits were assessed using the Neuroticism-Extraversion-Openness Five-Factor-Inventory (NEO-FFI[23]). Past month use of soft and hard drugs was assessed with a self-report questionnaire. Alcohol use was assessed with the alcohol use disorders identification test (AUDIT[24]). Medication was classified according to the World Health Organization Anatomical Therapeutic Chemical classification[25]. Frequent use (daily or at least 50% of the time) of tricyclic antidepressants (TCA; N06AA), selective serotonin reuptake inhibitors (SSRI; N06AB), and other antidepressants (including monoamine oxidase inhibitors, nonselective [N06AF], and antidepressants classified as N06AX) as well as benzodiazepines (N03AE, N05BA, N05CD, and N05CF) was assessed, based on drug container inspection during the baseline interview. The presence of chronic somatic illness was assessed during the face-to-face interview, a count was made of both the number of present chronic illnesses and the number of chronic illnesses under treatment. Inflammation was also assessed as it reflects poor somatic health and is considered to be an important risk factor for depression[26]. Levels of inflammation markers C-reactive protein (CRP), Interleukin-6 (IL-6), and tumor necrosis factor alpha (TNF-a) were determined in blood samples, obtained on the day of the interview around 8:00 AM[27]. The inflammation markers were log transformed prior to analysis.

STATISTICAL ANALYSES

Missing data

In those with five or less missing responses, missing IDS-SR item scores were imputed for baseline (323 item scores, 0.50%) and follow-up data (133 item scores, 0.24%). Missing values on the external variables were imputed on scale level for baseline (1419 scale scores, 0.96%) and follow up data (684 scale scores, 0.35%). Imputation was performed using the R package 'impute'[28] using a K-nearest neighbor (KNN) search to impute missing values based on scores of subjects with similar symptom profiles. KNN imputation was chosen based on theoretical grounds because differences in symptom reporting were hypothesized to exist across the sample and an imputation approach based on the whole sample would be in contradiction with this.

Exploratory analyses

Prior to the person-fit analyses, exploratory nonparametric IRT analyses were performed to inspect data quality and check IRT assumptions. Nonparametric IRT analyses provide insight in the quality of data and suitability of the data for parametric modelling[29].

Mokken scale analysis was performed to inspect IRT assumptions using MSP5.0[30] and 'mokken' R package[31]. Item response behavior was visually inspected using Testgraf[32]. Unidimensionality is a key assumption of IRT. The probability of reporting a symptom should mainly depend on the underlying level of depression and not on other, smaller factors. The extent of unidimensionality was checked by fitting a bi-factor model with each item loading on a general factor and on a specific group factor. Previous factor analytic studies on the IDS-SR[19,33] were used to specify the factor structure. The root mean square error of approximation (RMSEA) and the comparative fit index (CFI) were used to assess model fit, with a RMSEA<0.06 and CFI>0.95 indicating good model fit[34]. Two criteria were used to assess the unidimensionality of the data: (a) factor loadings on the general factor and the group factors were compared[35], and (b) the explained common variance (ECV) of the general factors was evaluated[36]. These analyses were performed with the R-package 'lavaan'[37] using adjusted weighted least squares (WLSMV) estimation.

Person-fit analyses

The graded response model (GRM[38]) was fitted as IRT model. The GRM was chosen because the item response data consisted of ordered categorical responses reflecting symptom severity[39]. The model estimates two parameters for each item: the *discrimination parameter (α)* describes how strong a symptom (item) is related to underlying depression severity (person characteristic), and the *threshold (β)* reflects the severity of a symptom (item).

Person-fit analysis using probability based IRT models enable identification of individuals that respond in a different way than would be expected based upon the IRT model that is used to describe the data[40]. The probability of endorsing depressive symptoms is expected to decrease as symptoms become more severe. In many cases, misfit is caused by the endorsement of severe symptoms (e.g. suicidal ideation) without endorsement of milder symptoms (e.g. sad mood). The more a person's responses deviate from this pattern of decreasing endorsement with increasing severity, the poorer his/her person-fit. In the current study, the polytomous likelihood based person-fit statistic l_z[41]was used on baseline and follow-up data. The l_z index represents how likely it is to observe an individual's pattern of endorsed items, given the estimated IRT model. Persons with response patterns that are consistent with the IRT model will have higher values of l_z, indicating good person fit, whereas persons with atypical response patterns will have low values of l_z, indicating poor person fit.

Participants were divided into three groups based on their person-fit score: an 'atypical' group identified as participants with poor person-fit scores below a 5% significance level, a 'prototypical' group with excellent person-fit scores above a 95% significance level, and a 'middle' group containing all other participants. Because the observed person-fit

was not normally distributed, the 5% and 95% significance level cut-offs were obtained from the values corresponding to the top and bottom 5% of an empirical distribution of the person-fit statistic using Monte Carlo simulation (10,000 simulees) with the sample characteristics of the real data taken into account.

Regularized regression

Because there was a large number of strongly inter-correlated external variables, conventional multivariate regression methods were inappropriate to investigate external correlates of atypical response patterns. Therefore, a shrinkage and variable selection technique called 'elastic net' regression was used instead[42]. Elastic net is a data-mining method that can be used to fit a prediction model with a very large number of predictors. A penalty is applied to the coefficients of the fitted model to shrink the coefficients of predictors with small effects to zero and to reduce their variance. In this way, all irrelevant effects are shrunk to zero and the most important predictors are selected from the initial large set of predictors. Fivefold cross validation across randomly selected subsamples was performed varying the tuning parameter (alpha) that controls the strength of the penalty (between 0=low penalty and 1=high penalty) to find the shrinkage that yielded the lowest cross-validated error, and thus, the most accurate cross-validated model. The selected regression coefficients were exponentiated to create odds-ratios.

RESULTS

DESCRIPTIVES

Baseline sample descriptives are shown in Table 1. The mean age was 42.2 (range 18-65), and 67.9% was female. Of the patients, 51.4% had a CIDI diagnosis of a mood disorder (MDD or dysthymia) and 55.9% of any anxiety disorder. At two-year follow-up (n=1942), the mean age was 44.4 (range 19-68) and 68.0% was female. In addition, the rates of CIDI mood disorders (30.0%) and anxiety disorders (34.2%) were lower than at baseline.

EXPLORATORY ANALYSES

Nonparametric IRT analyses showed violations of model assumptions caused by rank-order problems. For some items, the higher categories were equally or more often endorsed than lower categories. Item functioning improved when items were recoded from four to three categories. After this rescoring, unsatisfactory response behavior was still observed for items 1-4 ('onset insomnia', 'mid insomnia', 'morning insomnia' and 'hypersomnia') and item 9 ('mood variation'). Therefore, these items were excluded from further analyses.

TABLE 1. Descriptive characteristics of baseline sample (n=2292).

Characteristic	N (%)	Mean (SD)
Demographics		
Female gender	1557 (67.9%)	
North European ancestry	2167 (94.5%)	
Secondary care	975 (42.5%)	
Age (years)		42.2 (12.6)
Years of education		12.0 (3.3)
Personality		
Neuroticism		38.9 (8.1)
Extraversion		35.4 (7.0)
Openness		38.3 (6.0)
Agreeableness		43.3 (5.3)
Conscientiousness		40.8 (6.4)
Severity		
4DSQ – Somatization		13.6 (9.7)
4DSQ – Distress		35.3 (14.1)
Audit score		4.8 (5.0)
BAI score		14.4 (10.7)
IDS score[1]		25.2 (13.4)
MDQ score		5.2 (3.2)
Alcohol disorders[2]		
Alcohol abuse	264 (11.5%)	
Alcohol dependence	418 (18.2%)	
Mood disorders[2]		
MDD	1093 (47.7%)	
Dysthymia	299 (13.0%)	
Any	1179 (51.4%)	
Anxiety disorders[2]		
Social phobia	656 (28.6%)	
GAD	452 (19.7%)	
Panic disorder	656 (28.6%)	
Agoraphobia	183 (8.0%)	
Any	1282 (55.9%)	
Medication use		
Benzodiazepines	437 (19.1%)	
SSRI	498 (21.7%)	
TCA	77 (3.4%)	
Other antidepressants	164 (7.2%)	
Inflammation markers[3]		
C-reactive protein (mg/l)		1.31 (0.56-3.11)
Interleukin-6 (pg/ml)		0.78 (0.50-1.28)
Tumor necrosis factor-alpha (pg/ml)		0.80 (0.60-1.10)

SD, standard deviation.
[1] IDS sum score of original non-rescored data.
[2] CIDI diagnosis of current DSM-IV disorders (during past six months).
[3] Median and interquartile range of non-transformed inflammation markers.

The bi-factor model showed good fit to the data (CFI=0.99; RMSEA=0.03), with a strong general factor of 0.35 (range=0.19-0.50), considerable smaller loadings on the group-factors (average=0.10; range=-0.12-0.32) and an explained common variance (ECV) of 0.84. This indicated that most covariance was explained by the general factor and the data was sufficiently unidimensional for IRT analyses.

PERSON-FIT

The distribution of the person-fit scores (Figure 1) was skewed to the left, but had quite a uniform shape across different levels of depression severity, with slightly poorer person-fit with increasing depression severity (r=-0.13; 95% CI [-0.16,-0.08]).

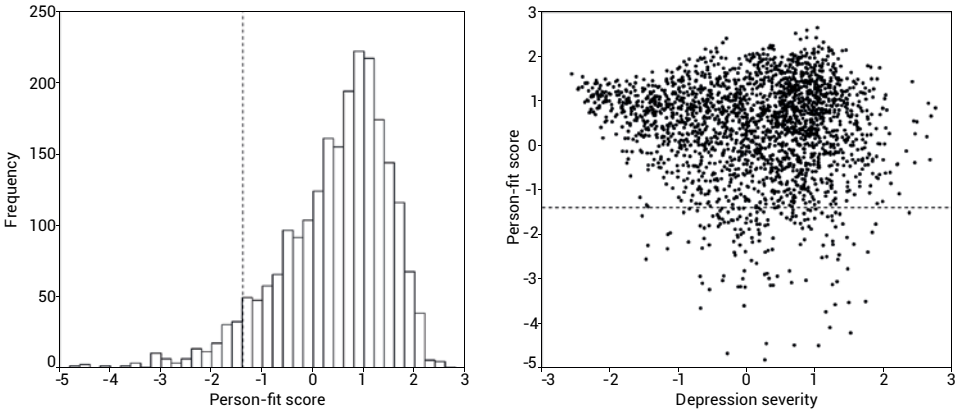


FIGURE 1. Distribution of person-fit scores, frequencies (left) and across different levels of depression severity (right) with the dotted line representing the empirically derived 5% person-fit cutoff (l_z=-1.39) for the atypical group.

To investigate what low person-fit means in practice, the response patterns and overall characteristics of the seven persons with the lowest person-fit scores (l_z<-4) are described in detail in Table 2. Inspection of these cases showed that their response patterns were characterized by (a) extreme scores and (b) reporting of severe symptoms but no milder symptoms. For example, patient 1145 with poor person-fit reported the following atypical symptom pattern (items ordered by increasing severity): 00022122200120220002000. This person often endorsed either the lowest (0) or highest (2) category and endorsed severe symptoms without endorsing milder symptoms, which is not in line with the underlying depression construct. In contrast, the person with the highest person fit (l_z=2.7) showed the following response pattern: 22111111111111111110100. This person gave less extreme responses and did report all symptoms that were milder than the most severe

endorsed symptoms, in line with the underlying construct. Taken together, these results illustrate how (low) person-fit is indicative of unexpected atypical response behavior on the IDS-SR.

TABLE 2. Example cases of patients with atypical profiles of depressive symptoms as identified by poor person-fit scores. The seven patients with person-fit scores on l_z below -4 are reported.

Patient	l_z	Age	Gender	IDS	Description
568	-4.1	38	Female	42	Reports many depressive symptoms including severe symptoms like 'suicidal thoughts' but does not report core symptoms 'sad mood' and 'capacity of pleasure'. Has a high IDS total score (42) indicative for severe depression, but has no CIDI diagnosis of MDD or dysthymia (current nor lifetime). The patient has three CIDI anxiety diagnoses past month (social phobia, panic with agoraphobia and generalized anxiety disorder) and scores high on the BAI (61).
1974	-4.2	58	Female	36	Scores low on mild symptoms, but high on severe symptoms 'suicidal thoughts' and 'weight gain'. Has a recurrent MDD with melancholic features, reported to have 4 previous episodes of MDD and is using tricyclic antidepressants.
592	-4.4	42	Female	19	Reports 0 on most items including 'sad mood' but scores high on somatic related symptoms. Has a high score of 18 on the 4DSQ somatization subscale and reports four chronic illnesses of which two under treatment.
935	-4.5	45	Female	32	Scores low on most symptoms including 'sad mood' but scores high on 'reactivity of mood', 'quality of mood', and 'suicidal thoughts'. Is diagnosed with MDD, social phobia, panic with agoraphobia, generalized anxiety disorder past month and CIDI alcohol disorder.
1145	-4.5	41	Male	31	The patient shows a pattern of extreme scores with high scores on all mood and anxiety related symptoms and low scores on all somatic related symptoms. The patient is diagnosed with a recurrent MDD and reports to have had 50 MDD episodes during lifetime.
61	-4.7	42	Male	17	Reports only high on a few symptoms including 'sad mood', 'reactivity of mood', 'quality of mood' and 'weight change'. Has a low IDS sum score of 17 indicating no depression, but has a CIDI diagnosis of MDD past month.
2936	-4.8	25	Male	26	Patient reports extreme with 0 on most symptoms and only high on a few symptoms 'sad mood', 'quality of mood', 'reactivity of mood', and 'panic'. Patient is diagnosed CIDI MDD with melancholic features and dysthymia past month.

PERSON-FIT SUBGROUPS

An atypical group (n=156, 6.8%) was identified with person-fit scores below the empirically derived 5% person-fit cutoff (l_z=-1.39), and a prototypical group (n=332, 14.4%) with person-fit scores above the 95% person-fit cutoff (l_z=1.47). The middle group (l_z=-1.39 and 1.47) consisted of 1804 persons.

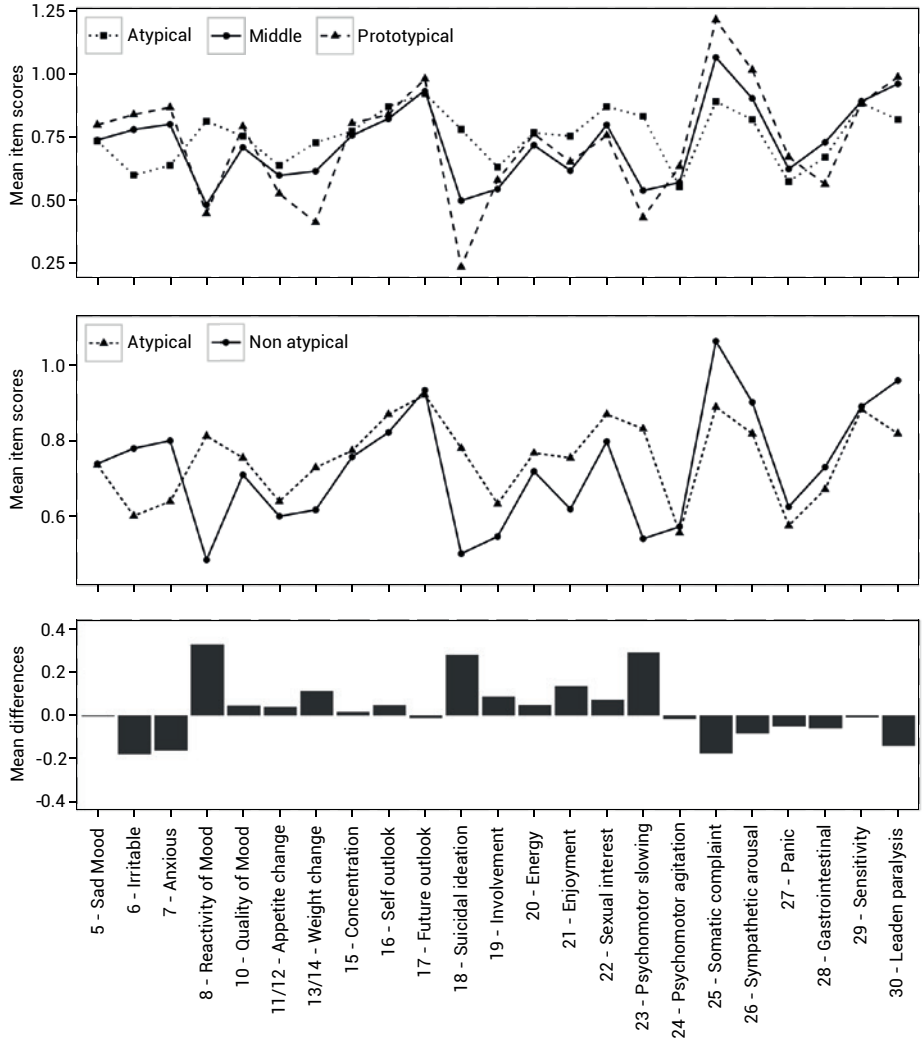


FIGURE 2. Differences in symptom profiles between the atypical, middle and prototypical group. The groups are defined based on their person-fit scores: the atypical group has poor person-fit (l_z<-1.39), the prototypical group has excellent person-fit (l_z>1.47) and the middle group has scores in between (-1.39<l_z<1.47). The top panel shows mean item scores for the atypical, middle and prototypical group. The middle and lower panel show mean item scores and mean differences for the atypical and non-atypical group (middle and prototypical combined).

Mean item scores of the atypical group differed substantially from the middle and prototypical group (Figure 2). Persons with atypical response patterns had lower mean scores on the items 'irritable', 'anxious', 'somatic complaint', and 'leaden paralysis' and considerably higher mean scores on 'reactivity of mood', 'weight change', 'suicidal

ideation', and 'psychomotor slowing'. Mean differences on core depression symptoms like 'sad mood', 'involvement' and 'concentration' were minimal. Interestingly, the atypical groups' mean item scores on 'reactivity of mood' (0.81) and 'suicidal ideation' (0.78) were even higher than the mean score on the core symptom 'sad mood' (0.75). Mean differences between the atypical and prototypical groups were larger than between the atypical and middle group, but showed similar qualitative patterns of item-score differences.

EXTERNAL VARIABLES

The regularized regression results are shown in Table 3. Linear elastic-net regression with continuous person-fit as outcome selected 17 out of 52 variables into a prediction model (regression coefficients of -0.27 to 0.37). Logistic elastic net regression with atypical group membership as outcome selected 16 variables into a prediction model (OR=0.6-1.5). When using prototypical group membership as outcome, 17 variables (OR=0.5-1.5) were selected. Predictors selected in all three models that were associated with atypical group membership, included the presence of melancholic features (OR=1.2) and atypical features (OR=1.2), first MDD onset longer than a year ago (OR=1.1), and medication use (OR 1.1-1.5). The predictors North-European ancestry (OR=0.6), female gender (OR=0.8), somatization (OR=0.8), anxiety severity (OR=0.8) and anxiety diagnoses (OR 0.8-0.9) showed negative associations with atypical group-membership in all models. Interestingly, TNF-α was positively associated with both atypical profiles (OR=1.1) and prototypical profiles (OR=1.1). Together, these results indicated that the extent of atypical response behavior on the IDS-SR is predicted by a combination of socio-demographic, psychiatric and biological external characteristics.

PERSON-FIT CONSISTENCY ON FOLLOW-UP

Person-fit at baseline and follow-up correlated positively (r=0.36). In addition, the empirically derived cutoffs were comparable between baseline and follow-up for atypical (l_z=-1.37) group membership, and slightly lower at follow-up for prototypical (l_z=1.36) group membership. The identified atypical group was slightly larger at follow-up (n=158, 8.0%) and the prototypical group was smaller (n=138, 7.1%). From the 156 patients that were in the atypical group at baseline, 126 (80.7%) provided data at follow up. Of these, 37 (29%) were again in the atypical group at follow-up. The qualitative patterns of differences between the mean item scores in the atypical, middle and prototypical groups at follow-up were comparable to those at baseline. Again, persons in the atypical group scored considerably lower than the middle and prototypical group on 'irritable', 'anxious' and considerably higher on 'mood reactivity' and 'suicidal ideation'.

TABLE 3. Elastic net coefficients[1] predicting atypical profiles of depressive symptoms based on person-fit statistic l_z.

		Linear[2]	Logistic[3]	
		Atypicality	Atypical	Prototypical
IDS	Total score[4]	0.02	1.0	1.0
	Melancholic features	0.04	1.2	0.8
	Atypical features	0.21	1.2	0.5
Demographics	Female gender	-0.07	0.8	1.1
	Healthcare setting	0.05	1.1	
	North-European ancestry	-0.27	0.6	1.2
Chronic illness	Chronic illness (total)	0.03	1.1	
	Chronic illness (under treatment)		0.9	
4DSQ	Somatization score ≥11	-0.11	0.8	1.3
BAI	Anxiety score ≥8	-0.26	0.8	1.5
MDQ	MDQ Bipolar ≥7		1.1	
Alcohol disorders	Alcohol dependence			0.9
	Alcohol abuse		1.1	
Mood disorders	Dysthymia (current)			0.9
	MDD (lifetime)	-0.15	0.8	1.2
	Recurrent MDD			1.1
	Onset first MDD episode (>12M)	0.10	1.1	0.8
Anxiety disorders	Agoraphobia (<6M)		0.8	
	Social phobia (lifetime)	-0.06	0.9	1.1
	Panic disorder (<6M)		0.9	0.8
	Panic disorder (lifetime)	-0.06		
Medication use	Benzodiazepines		1.2	
	TCA	0.37	1.5	0.6
	SSRI	0.03	1.1	0.9
	Other Antidepressant	0.15	1.2	
Biological	TNF-α			

[1] Elastic net regression models based on tuning and shrinkage parameter that gave the lowest five-fold cross validated error. Not all investigated predictors were selected by the models (e.g. age, education level, personality, inflammation markers CRP and IL-6 were not selected in any model). No confidence intervals are reported due to the bias in standard errors produced by these models.

[2] Linear regression with l_z person-fit negated as outcome variable, such that positive coefficients indicate the prediction of atypical response patterns.

[3] Logistic regression for groups defined by comparing person-fit scores to the empirical derived cutoffs for atypical ($l_z < -1.39$) and prototypical ($l_z > 1.46$). Odds-ratios are reported.

[4] Total score on IDS was included in all models to account for potential quantitative differences in depression severity across groups.

DISCUSSION

This study aimed to identify and investigate patients with unexpected, atypical profiles of reported depressive symptoms using a data-driven approach based on person-fit statistics. Using person-fit, a group of atypical responders on the IDS-SR was identified. Inspection of individual cases and their characteristics showed different types of atypical response profiles, each with unique descriptions and potential explanations for the atypical responding. Comparison of the item score patterns between the atypical, middle, and prototypical groups showed clear qualitative differences, with the atypical group reporting high on severe symptoms (e.g. 'mood reactivity' and 'suicidal ideation') while scoring low on mild symptoms (e.g. 'sad mood'). Additional analyses showed that atypical symptom profiles were associated with socio-demographic (e.g. gender), clinical (e.g. medication use) and biological (e.g. TNF-α) factors. Person-fit scores at follow-up were positively correlated with person-fit scores at baseline and a considerable number of atypical responders at baseline were also in the atypical group at follow-up.

Case descriptions of the seven patients with poorest person fit confirmed that low person-fit scores are indicative of a variety of depressive symptom profiles that do not adhere to the underlying depression severity model. Several response-behavior characteristics could be noted from these case descriptions. First, the item-scores suggested that IDS total scores were not likely to reflect depression severity, limiting clinical interpretability. For example, one patient reported a high IDS sum score of 42, indicative of depression, but had no CIDI MDD diagnosis. Conversely, another patient had a low IDS sum score of 17 but also a current MDD diagnosis (past month) according to the CIDI. Second, for each atypical case, possible explanations could be found for the observed inconsistencies in symptom profiles. However, these explanations differed across patients and were specific to particular response patterns. The results indicate that external factors are associated with atypical depressive symptom reporting and that identification of patients with such profiles could help when trying to distinguish between those who report symptoms in the context of depression and those who do so for other reasons. Third, several of the investigated atypical cases showed symptom profiles with extreme scores and endorsed severe symptoms (e.g. mood reactivity, suicidal ideation) without endorsing milder symptoms (e.g. sad mood, capacity of pleasure).

Although the individual cases with the poorest person-fit showed very distinct case descriptions, group-level comparisons of item score patterns did reveal clear differences between the atypical and the other subgroups. The strongest differences were observed for items 'mood reactivity' and 'suicidal ideation', which were more often endorsed in the atypical group. This pattern could be explained in different ways. Mood reactivity in the absence of milder symptoms could be an early sign of depression relapse/remission[43].

Reporting suicidal ideation in the absence of milder symptoms fits in with previous findings about atypical suicide risk[17], which is considered a subtype of suicide risk where patients report suicidal ideation in isolation from other depressive symptoms (e.g. no sad mood). Detection of patients at risk of relapse and/or suicide risk is important and the person-fit statistic could be a useful tool to do this.

Patients in the atypical group scored lower on items 'anxiety', 'irritability', and 'somatic complaint'. Also, anxiety diagnoses, BAI anxiety score, and 4DSQ somatization score were negatively associated with atypical response behavior. This could be explained by the fact that symptoms of anxiety and somatization are frequently observed in patients with MDD, and therefore obtain lower thresholds in the IRT model[44]. For most patients, reporting mild anxiety or somatization merely results in a slightly higher score on the underlying latent trait (depression severity), and does generally not result in atypical profiles. However, in some cases (e.g. patient 568) severe anxiety or somatization can result in atypical profiles when severe symptoms (e.g. suicidal ideation) but no mild depressive symptoms (e.g. sad mood) are reported.

The use of psychoactive medication was found to be associated with atypical depressive symptom profiles. This could be explained by the asymmetrical effects of medication on depressive symptoms. For instance, antidepressants may primarily elevate mood, with improvement of other symptoms as a secondary consequence[45]. As a result, a patient may report improvement on milder symptoms but less so on severe symptoms, which respond slower to medication. Also, the strong association of person fit with TCAs may be explained by the fact that patients who are prescribed TCAs are usually complex and difficult to treat cases[46].

Interestingly, the inflammation marker TNFα was predictive of both atypical and prototypical profiles. The effect is small but could be explained by the fact that heightened levels of inflammation are on the one hand associated with antidepressant use[27], which was found to be associated with atypical responding, and on the other hand with somatization[26], which was found to be associated with prototypical responding.

The person-fit statistic showed considerable consistency over time. Person-fit scores were positively correlated between baseline and two-year follow-up, the item-score patterns were similar and a substantial part of the baseline atypical subgroup also showed an atypical profile at follow-up. This consistency of person-fit is remarkable given the fact that it is merely a statistic to quantify deviations from an IRT model and the considerable follow-up period. The finding that person-fit is consistent for a considerable number of patients could indicate that their atypical reporting is related to a trait-like pathology that consistently produces atypical symptom profiles[16].

This study has several strengths, including the large sample size, the use of advanced psychometric techniques, the use of a large range of external variables, and the possibility to investigate consistency over time. However, there were also limitations. First, the study excluded several psychopathological conditions (e.g. bipolar disorder[47]). Second, the results are based on the IDS-SR and their generalizability to other instruments needs to be evaluated. Third, the results apply to depressive outpatients with relatively mild symptomatology and cannot be generalized to other populations (e.g. depressive inpatients). Fourth, although the external associates of atypical symptom reporting were thoroughly investigated, these group-based results only tell part of the story. The case descriptions of the patients with the poorest person-fit scores showed that each person had unique characteristics, illustrating how heterogeneous the group of atypical responders actually is and suggesting that variation in response behavior can only be partly explained on the group-level. This could be problematic for the investigation of potential causes of atypical responding. However, it is important to realize that irrespective of the cause of atypical responding, patients identified with atypical response patterns have total scores that are not a valid reflection of depression severity. Finally, the current study is a methodological study. Person-fit statistics seem promising and potentially clinically relevant, but a clinical field study is needed to further evaluate its practicality and acceptability in practice.

In conclusion, a data-driven approach was shown to be useful for the identification of patients who report atypical profiles of depressive symptoms on the IDS-SR. Person-fit statistics allowed for a novel and interesting approach to investigate the heterogeneity of depression by decomposing the sample into those with typical and atypical response profiles. Results show that a range of factors influence the reporting of depressive symptoms, which can lead to atypical profiles that do not conform to models of depression, and cannot be scaled or classified accordingly. When using depression questionnaires to assess their patients, clinicians could potentially use person-fit statistics to detect atypical responding. For example, person-fit could be computed for each patient who completes a computer-administered depression questionnaire. A warning could be given if the reported symptom pattern is potentially atypical, given the underlying depression construct. Clinicians can then closer inspect the item scores and augment the assessment with additional information about the patient to judge whether a depression scale score is a valid reflection of underlying symptomatology.

REFERENCES

1. Lux, V. & Kendler, K. Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria., Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychol Med* **40, 40,** 1679, 1679–1690 (2010).
2. Widiger, T. A. & Clark, L. A. Toward DSM-V and the classification of psychopathology. *Psychol Bull* **126,** 946–963 (2000).
3. Widiger, T. A. & Samuel, D. B. Diagnostic categories or dimensions? A question for the Diagnostic and statistical manual of mental disorders–fifth edition. *Journal of Abnormal Psychology* **114,** 494–504 (2005).
4. Penninx, B. W. J. H. *et al.* Two-year course of depressive and anxiety disorders: Results from the Netherlands Study of Depression and Anxiety (NESDA). *Journal of Affective Disorders* **133,** 76–85 (2011).
5. Wardenaar, K. J., Conradi, H.-J. & de Jonge, P. Data-Driven Course Trajectories in Primary Care Patients with Major Depressive Disorder. *Depress Anxiety* **31,** 778–786 (2014).
6. Luyten, P., Blatt, S. J., Van Houdenhove, B. & Corveleyn, J. Depression research and treatment: Are we skating to where the puck is going to be? *Clinical Psychology Review* **26,** 985–999 (2006).
7. Stewart, J. W., McGrath, P. J., Quitkin, F. M. & Klein, D. F. Atypical depression: current status and relevance to melancholia. *Acta Psychiatrica Scandinavica* **115,** 58–71 (2007).
8. Baumeister, H. & Parker, G. Meta-review of depressive subtyping models. *Journal of Affective Disorders* **139,** 126–140 (2012).
9. Sullivan, P. F., Kessler, R. C. & Kendler, K. S. Latent Class Analysis of Lifetime Depressive Symptoms in the National Comorbidity Survey. *AJP* **155,** 1398–1406 (1998).
10. van Loo, H. M., de Jonge, P., Romeijn, J.-W., Kessler, R. C. & Schoevers, R. A. Data-driven subtypes of major depressive disorder: a systematic review. *BMC Medicine* **10,** 156 (2012).
11. Lubke, G. H. & Muthén, B. Investigating Population Heterogeneity With Factor Mixture Models. *Psychological Methods* **10,** 21–39 (2005).
12. Leentjens, A. F. G., Verhey, F. R. J., Luijckx, G.-J. & Troost, J. The validity of the Beck Depression Inventory as a screening and diagnostic instrument for depression in patients with Parkinson's disease. *Mov. Disord.* **15,** 1221–1224 (2000).
13. Embretson, S. & Reise, S. *Item Response Theory for Psychologists*. (Psychology Press, 2000).
14. Meijer, R. R. Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods* **8,** 72–87 (2003).
15. Conijn, J. Detecting and explaining person misfit in non-cognitive measurement. (2013).
16. Woods, C. M., Oltmanns, T. F. & Turkheimer, E. Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment* **20,** 159–168 (2008).
17. Conrad, K. J. *et al.* Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence* **106,** 92–100 (2010).
18. Penninx, B. W. J. H. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17,** 121–140 (2008).
19. Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B. & Trivedi, M. H. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychological Medicine* **26,** 477–486 (1996).
20. Beck, A. T., Epstein, N., Brown, G. & Steer, R. A. An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology* **56,** 893–897 (1988).
21. Hirschfeld, R. M. A. *et al.* Development and Validation of a Screening Instrument for Bipolar Spectrum Disorder: The Mood Disorder Questionnaire. *AJP* **157,** 1873–1875 (2000).
22. Terluin, B. *et al.* The Four-Dimensional Symptom Questionnaire (4DSQ): a validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. *BMC Psychiatry* **6,** 34 (2006).
23. Costa Jr, P. & McCrae, R. *Neo personality inventory–revised (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual.* (Psychological Assessment Resources, 1992).
24. Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R. & Grant, M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-II. *Addiction* **88,** 791–804 (1993).

25. World Health Organization Collaborating Centre for Drug Statistics Methodology. *Anatomical Therapeutic Chemical (ATC) classification*. (World Health Organization, 2007).
26. Dantzer, R., O'Connor, J. C., Freund, G. G., Johnson, R. W. & Kelley, K. W. From inflammation to sickness and depression: when the immune system subjugates the brain. *Nat Rev Neurosci* **9,** 46–56 (2008).
27. Vogelzangs, N. *et al.* Association of depressive disorders, depression characteristics and antidepressant medication with inflammation. *Transl Psychiatry* **2,** e79 (2012).
28. Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. *impute: Imputation for microarray data.* (2013).
29. Meijer, R., Tendeiro, J. & Wanders, R. in *Handbook of item response theory modeling: Applications to typical performance assessment* 85–110 (Routledge, 2014).
30. Molenaar, I. & Sijtsma, K. *User's manuals MSP5 for Windows. IEC ProGAMMA, Groningen.* (2000).
31. van der Ark, L. Mokken scale analysis in R. *Journal of Statistical Software* **20,** (2007).
32. Ramsay, J. *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*. (2000).
33. Wardenaar, K. J. *et al.* The structure and dimensionality of the Inventory of Depressive Symptomatology Self Report (IDS-SR) in patients with depressive disorders and healthy controls. *Journal of Affective Disorders* **125,** 146–154 (2010).
34. Hu, L. & Bentler, P. M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* **6,** 1–55 (1999).
35. McDonald, R. *Test theory a unified treatment*. (L. Erlbaum Associates, Mahwah, N.J., 1999).
36. Reise, S. P., Moore, T. M. & Haviland, M. G. Bifactor Models and Rotations: Exploring the Extent to Which Multidimensional Data Yield Univocal Scale Scores. *Journal of Personality Assessment* **92,** 544–559 (2010).
37. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* **048,** (2012).
38. Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement* **34,** 100 (1969).
39. Reise, S. P. & Waller, N. G. How many IRT parameters does it take to model psychopathology items? *Psychological Methods* **8,** 164–184 (2003).
40. Meijer, R. R. & Sijtsma, K. Methodology Review: Evaluating Person Fit. *Applied Psychological Measurement* **25,** 107–135 (2001).
41. Drasgow, F., Levine, M. V. & Williams, E. A. Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology* **38,** 67–86 (1985).
42. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301–320 (2005).
43. Segal, Z. V. *et al.* Cognitive Reactivity to Sad Mood Provocation and the Prediction of Depressive Relapse. *Arch Gen Psychiatry* **63,** 749–755 (2006).
44. Wanders, R. B. K. *et al.* Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make? *Journal of Psychosomatic Research* **78,** 130–136 (2015).
45. Harmer, C. J., Goodwin, G. M. & Cowen, P. J. Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *The British Journal of Psychiatry* **195,** 102–108 (2009).
46. Rosholm, J.-U., Andersen, M. & Gram, L. F. Are there differences in the use of selective serotonin reuptake inhibitors and tricyclic antidepressants? A prescription database study. *Eur J Clin Pharmacol* **56,** 923–929 (2001).
47. Zimmermann, P. *et al.* Heterogeneity of DSM-IV Major Depressive Disorder as a Consequence of Subthreshold Bipolarity. *Arch Gen Psychiatry* **66,** 1341–1352 (2009).

CHAPTER 3

What Does the Beck Depression Inventory Measure in Myocardial Infarction Patients? A Psychometric Approach Using Item Response Theory and Person-Fit

Klaas J. Wardenaar, Rob B.K. Wanders, Annelieke M. Roest, Rob R. Meijer, Peter de Jonge

International journal of methods in psychiatric research 2015, *24:130-142*.

ABSTRACT

Observed associations between depression following myocardial infarction (MI) and adverse cardiac outcomes could be overestimated due to patients' tendency to over report somatic depressive symptoms. This study was aimed to investigate this issue with modern psychometrics, using item response theory (IRT) and person-fit statistics to investigate if the Beck Depression Inventory (BDI) measures depression or something else among MI-patients.

An IRT-model was fit to BDI-data of 1135 MI patients. Patients' adherence to this IRT-model was investigated with person-fit statistics. Subgroups of 'atypical' (low person-fit) and 'prototypical' (high person-fit) responders were identified and compared in terms of item-response patterns, psychiatric diagnoses, socio-demographics and somatic factors. In the IRT model, somatic items had lower thresholds compared to depressive mood/cognition items. Empirically identified 'atypical' responders (n=113) had more depressive mood/cognitions, scored lower on somatic items and more often had a CIDI depressive diagnosis than 'prototypical' responders (n=147). Additionally, 'atypical' responders were younger and more likely to smoke. In conclusion, the BDI measures somatic symptoms in most MI patients, but measures depression in a subgroup of patients with atypical response patterns. The presented approach to account for interpersonal differences in item responding could help improve the validity of depression assessments in somatic patients.

INTRODUCTION

Research findings suggests that there is a relationship between acute coronary syndromes (ACS), such as myocardial infarctions (MI), and depression. In community samples, depression has been found to be a risk factor for cardiovascular disease (CVD[1]) and in ACS patients, depression has been found to increase the risk of new cardiac events[2] and mortality[2–6]. Conversely, MI has been found to increase the risk of depression[7] and depression chronicity[8]. However, the supposed bidirectional link between depression and ACS is controversial. An often-heard criticism is that the association may be overestimated due to biased depression measurements[9–12]. Widely used questionnaires such as the Beck Depression Inventory (BDI) include items that asses somatic/functioning symptoms, which are common in depression but also in somatic illness, potentially leading to overestimated depression scores in patients with a somatic illness[12–14].

Despite the importance of issues with depression measurement in psychosomatic research, thorough investigations with modern psychometrics have been scarce. Previous factor analytical studies in somatic patients have shown that somatic items are part of a general depression severity scale but also constitute a specific domain within depression questionnaires[11,15]. This suggests that the endorsement of somatic items is not only explained by the presence of depression, but also by other sources, such as somatic problems and/or functional impairments. Although abovementioned studies have shown the underlying structure of depression-related symptoms, they provide no insight into the endorsement probabilities of individual items. This makes factor analyses of limited use when trying to understand depressive symptom reporting in somatically ill patients. Investigation of the latter with Item Response Theory (IRT [16]) could help to better understand how each symptom contributes to somatic patients' observed depression scores.

In an IRT model, each item in a questionnaire has a threshold on an underlying severity dimension that reflects its endorsement probability. Items that assess mild symptoms have low thresholds and are frequently reported; severe items have high thresholds and are less frequently reported. In an IRT model, the items' thresholds and slopes are estimated based on raw item response data, providing insight into the measurement characteristics of the instrument. In somatic patients, somatic items of a depression scale (e.g. 'energy loss') may be more frequently reported and thus have lower thresholds compared to items that cover depressive mood and cognitions (e.g. 'sadness', 'feeling guilty' [17]). This would entail that depression scores in this group are more reflective of somatic problems/functional impairments than of the full breadth of depressive symptomatology.

However, within a sample of patients there is still considerable interpersonal variation in the level of adherence to the group-based IRT model that is estimated based on all

subjects' data. In an MI sample, for example, the group-based IRT model may suggest that somatic problems are over-reported. In such a case, depressed patients that mainly report depressive mood/cognitions will not obey to the group-based IRT model and show a response pattern that is atypical for the group. This depressed subgroup is very interesting for psychosomatic research, but is easily overlooked when, based on the group-based IRT findings, all patients with increased BDI (or other questionnaire) scores are dismissed as merely over-reporting somatic depressive symptoms. A person-centered IRT approach can help to identify persons that do not conform to the group-based IRT model. In this approach, each person's level of adherence to the group-based IRT-model can be expressed by a person-fit statistic[18,19]. High person-fit indicates strong adherence to the group IRT-model (i.e. typical response patterns) and low person-fit indicates poor adherence (i.e. atypical response patterns). As such, person-fit can be used to investigate the interpersonal variations in item reporting and the factors that influence what a questionnaire measures across different persons. Eventually, this approach could help to better distinguish those patients, for whom increased depression scores reflect depression from those patients, for whom increased scores reflect only somatic illness.

The current study aimed to use a combination of group-level IRT and person-fit in a large MI patient sample (n=1135) to identify and investigate patients with (somatically) biased BDI scores and patients with BDI scores reflecting true depression. The analyses were conducted in several steps. First, an IRT-model was fit to the complete BDI-data to investigate the item-characteristics. Second, each patient's person-fit was calculated and finally the association of person-fit with interview-based DSM-IV major depressive disorder (MDD) and other external variables was investigated.

METHODS

PARTICIPANTS AND PROCEDURES

Data came from two studies with similar inclusion criteria that were combined in previous studies as well[20,21]. The Myocardial Infarction and Depression Intervention Trial (MIND-IT) study[22,23] was a multicenter randomized trial to investigate the effects of antidepressants in depressed MI patients. The Depression after Myocardial Infarction (DepreMI) study[24] was a naturalistic cohort study to investigate the effects of depression on cardiovascular outcome in MI patients.

In MIND-IT, MI-patients were recruited from 11 hospitals in the Netherlands. Inclusion criteria were: age ≥18 and a documented increase in cardiac enzymes together with at least 20 minutes of chest pain or electrocardiographic (ECG) changes typical of an MI. Exclusion

criteria were: the presence of disease influencing short-term life expectancy, inability to participate (e.g. communication problems, absence), receiving psychiatric treatment and/or participating in another trial. Of 2177 recruited patients, 331 met criteria of post-MI depression and were randomized to treatment (antidepressants and/or psychotherapy) or care-as-usual.

In DepreMI, 528 MI patients were recruited from 4 hospitals in the Netherlands. Patients were included if they had increased cardiac enzymes, chest pain for at least 20 minutes and pathological Q-waves in their ECG in at least two leads. Patients were excluded if they had a life expectancy <1 year (due to non-cardiac illnesses), were in poor physical condition, had cognitive problems, spoke insufficient Dutch, and/or were scheduled to have their future check-ups in a non-participating hospital. Both MIND-IT and DepreMI were approved by the institutional ethical review boards of their respective participating institutions. All patients provided informed consent.

From both samples, baseline BDI data collected before the administration of any treatment were combined in a single dataset. Only patients with a complete BDI and psychiatric interview (Comprehensive International Diagnostic Interview [CIDI] [25]) were included in the analyses to enable a comparison between the BDI and interview-based depression diagnoses. In DepreMI, the CIDI was administered to most patients (n=442) and to a subsample with a BDI≥10 in MIND-IT (n=760). Taken together, 1202 patients had the required data. Of these, 67 (5.6%) were excluded because they had missing responses on 5 or more BDI items. The study sample consisted of 1135 patients.

MEASURES

Depression

The BDI version 1[26,27] was administered to measure depression severity. The 21 items were scored on a 4-point Likert scale (0-3). In DepreMI, the CIDI 1.1[25] was administered and in MIND-IT, the CIDI 2.1[28] was administered to evaluate whether ICD-10 criteria[29] were met for depression after MI. The presence of anxiety disorders following MI (generalized anxiety disorder [GAD], panic disorder, social phobia, agoraphobia) was also assessed with the CIDI.

Demographic and clinical characteristics

All baseline clinical and demographic characteristics were assessed during hospitalization from the hospital charts. Left ventricular ejection fraction (LVEF) was assessed with radionuclide ventriculography, echocardiography, gated single photon emission computed tomography, angiography, magnetic resonance imaging, or clinical assessment. The following clinical variables were used in the current study: LVEF, Killip class, anterior site of

MI, history of MI, history of cerebral vascular disease, history of peripheral vascular disease, family history of coronary artery disease, diabetes, hypertension, hypercholesterolemia, body mass index (BMI) and current smoking.

Missing data

One-hundred-fifty-seven (13.8%) subjects had ≤3 missing responses, which were imputed with a k-nearest-neighbor (KNN) search[30]. KNN selects a number (k) of subjects that are most similar in terms of their response pattern to the subject, whose missing responses are to be imputed. Next, KNN imputes the weighted mean of the nearest neighbors' responses on the target item. Imputations were done with R-package 'impute'[31] with default settings (k=10). This method was chosen because item-reporting was hypothesized to vary across persons and imputation of scores calculated on the group level would not be in line with this.

STATISTICAL ANALYSES

Non-parametric item response theory

Preliminary non-parametric IRT analyses[32] were done with R-package 'KernSmoothIRT'[33] to evaluate the suitability of the data for parametric IRT modeling by plotting the non-parametric probability curves[34] of each item's categories. The plots were inspected to check whether item-responses were meaningfully related to underlying severity[35,36]. Items that showed response behavior that was not in line with IRT assumptions were removed from subsequent analyses.

Factor analysis

An exploratory factor analysis (EFA) was conducted with the BDI data, using a Weighted Least Squares (WLSMV) estimator for use with ordinal variables in Mplus (version 5 [37]). The ratio between the Eigenvalues of the first and the respective Eigenvalues of the second, third and fourth factor were inspected to evaluate the extent of unidimensionality[38].

Item response Theory

A graded response model (GRM [39]) was fit to the data with R-package 'ltm'[40]. Rather than choosing a model based on model fit, GRM was chosen a priori because it corresponded best with the categorical, ordered nature of the used data[41]. Two parameters were fit for each item: the slope indicates the strength of the relationship between the item and underlying severity and the thresholds (0-1, 1-2, 2-3) indicate the severity of the symptom that is assessed by the item. The items' IRT parameters were inspected and items were ordered by their average thresholds.

Person-fit

The likelihood-based person-fit statistic l_z[42] was used to quantify persons' adherence to the fitted IRT model. The l_z statistic reflects the likelihood of observing a response pattern, given the group-based IRT model. High l_z values (high person-fit) indicate strong consistency with the IRT-model whereas low l_z values indicate that a person's response pattern is atypical and adheres poorly to the group-based IRT-model[18,19]. To gain insight into its external correlates, person-fit was used as a continuous outcome variable in univariable linear regression analyses with external variables as determinants. Univariable analyses were conducted first, followed by a multivariable analysis including all independent variables with a univariate p-value <0.10. In another approach, external characteristics were compared between person-fit subgroups to gain insight into the more patient-specific characteristics of those with low vs. high person-fit. Because the person-fit statistic was not normally distributed, a Monte Carlo simulation procedure was used to simulate an empirical person-fit distribution with the same sample characteristics as the observed data, from which empirically-based unbiased person-fit cut-offs at a 10% significance level were derived[43]. The values corresponding to the 10% upper and lower part of the obtained empirical distribution were used as cutoffs to allocate patients in, respectively, prototypical and atypical subgroups based on their person-fit values. The other patients were allocated to a middle group. The atypical group (n=112, 9.8%) had person-fit below the 10% person-fit cut-off (l_z = -1.19) and the prototypical group (n=148, 13.0%) had person-fit above the 90% person-fit cut-off (l_z = 1.08). The 'middle' person-fit subgroup consisted of 874 patients.

RESULTS

DEMOGRAPHIC AND CLINICAL CHARACTERISTICS

The descriptives of the study sample are shown in the left-most column of Table 1. Of the sample, 76.9% was male and the mean age was 60.6 years (s.d.=11.8). Of the patients, 27.8% had a low LVEF (<45), 13.3% had a Killip class≥2 and 33.9% had an anterior site of the index MI. Several patients had a history of a prior MI (15.7%), CVD (5.7%) and/or PVD (9.0%). Health-related factors, such as current smoking (48.5%), hypercholesterolemia (61.1%), hypertension (32.2%), and family history of CVD (43.5%) were all common. The mean BDI score was 9.5 (s.d.=6.7) and 40.6% of the sample had a CIDI diagnosis of post-MI depression. The most prevalent post-MI anxiety diagnoses were GAD (8.2%), social phobia (3.3%) and agoraphobia (2.4%).

TABLE 1. Complete sample and subgroup characteristics in a group of patients with an acute coronary syndrome (N=1135)

		Total sample N=1135	Person-fit subgroups			Test-statistic	p
			'Atypical ' n=113 (10.0%)	'Average' n=874 (77.0%)	'Prototypical' n=148 (13.0%)		
Person-fit (l_z), range		-4.77-1.82	-4.77 to -1.20	-1.18 to 1.08	1.08 to 1.82	-	
Age mean years (SD)		60.6 (11.8)	57.9 (11.9)	60.7 (11.9)	62.3 (11.0)	F=4.6	0.01
Female gender, n (%)		262 (23.1%)	24 (21.2%)	208 (23.8%)	30 (20.3%)	X1.12=2	0.57
Cardiac severity	LVEF <45, n (%)	316 (27.8%)	32 (33.3%)	245 (34.5%)	39 (32.8%)	X0.17=2	0.92
	Killip class ≥2, n (%)	151 (13.3%)	13 (11.5%)	118 (13.5%)	20 (13.7%)	X0.37=2	0.83
	Anterior site of MI, n (%)	385 (33.9%)	42 (37.2%)	297 (34.0%)	46 (31.1%)	X1.07=2	0.59
Cardiac vulnerability	History of MI, n (%)	178 (15.7%)	18 (15.9%)	138 (15.8%)	22 (15.1%)	X0.06=2	0.97
	History of cerebral vascular disease, n (%)	66 (5.8%)	8 (7.1%)	55 (6.3%)	3 (2.1%)	X4.89=2*	0.08
	History of peripheral vascular disease, n (%)	102 (9.0%)	7 (6.2%)	84 (9.6%)	11 (7.6%)	X1.85=2	0.40
	Family history of coronary artery disease, n (%)	494 (43.5%)	49 (44.1%)	390 (45.0%)	55 (37.4%)	X2.96=2	0.23
Somatic health	Diabetes, n (%)	140 (12.3%)	16 (14.2%)	103 (11.8%)	21 (14.2%)	X1.05=2	0.59
	Hypertension, n (%)	366 (32.2%)	31 (27.4%)	285 (32.7%)	50 (33.8%)	X1.43=2	0.49
	Hypercholesterolemia, n (%)	694 (61.1%)	78 (69.0%)	523 (60.0%)	93 (62.8%)	X3.63=2	0.16
	Body mass index, mean (SD)	26.6 (4.0)	26.6 (4.4)	26.6 (3.9)	26.9 (4.7)	F=0.40	0.67
	Current smoking, n (%)	550 (48.5%)	71 (65.7%)	420 (49.6%)	59 (41.8%)	X14.5=2	0.001
CIDI depression after MI	Post-MI CIDI depression, n (%)	461 (40.6%)	58 (51.3%)	347 (39.7%)	56 (37.8%)	X6.15=2	0.046
	Number of depressive symptoms, mean (SD)	3.0 (2.8)	3.6 (3.0)	3.0 (2.8)	2.6 (2.7)	F=3.09	0.046
	Depressive mood, n (% present)**	450 (40.4%)	55 (50.0%)	339 (39.5%)	56 (38.6%)	X4.7=2	0.10
	Anhedonia/loss of interest, n (% present)**	371 (33.5%)	39 (35.1%)	285 (33.4%)	47 (32.6%)	X0.2=2	0.91

TABLE 1. Complete sample and subgroup characteristics in a group of patients with an acute coronary syndrome (N=1135) *(Continued)*

		Total sample N=1135	Person-fit subgroups				
			'Atypical ' n=113 (10.0%)	'Average' n=874 (77.0%)	'Prototypical' n=148 (13.0%)	Test-statistic	p
CIDI depression after MI	Energy loss, n (% present)**	503 (45.6%)	56 (51.4%)	388 (45.7%)	59 (40.7%)	X2.9=2	0.24
	Appetite change, n (% present)**	63 (5.7%)	9 (8.4%)	43 (5.1%)	11 (7.5%)	X3.0=2	0.22
	Sleeping problems, n (% present)**	529 (48.3%)	55 (52.4%)	411 (48.8%)	63 (42.9%)	X2.5=2	0.29
	Psychomotor agitation/retardation, n (% present)**	347 (31.2%)	41 (36.9%)	267 (31.2%)	39 (26.7%)	X3.1=2	0.22
	Feeling Worthless/Guilty, n (% present)**	229 (20.6%)	30 (27.8%)	179 (20.9%)	20 (13.8%)	X7.5=2	0.02
	Loss of Self-esteem, n (% present)**	218 (19.6%)	33 (30.8%)	160 (18.6%)	25 (17.2%)	X9.6=2	0.01
	Concentration problems, n (% present)**	455 (41.8%)	52 (48.1%)	350 (41.7%)	53 (37.3%)	X3.0=2	0.23
	Preoccupations with death/suicide, n (% present)**	217 (19.7%)	20 (20.0%)	167 (19.8%)	28 (19.0%)	X0.1=2	0.98
	Number of previous episodes, mean (SD)	0.63 (4.5)	1.4 (6.0)	0.55 (4.5)	0.50 (2.6)	F=2.0	0.14
CIDI anxiety after MI	Generalized Anxiety Disorder, n(%)	94 (8.3%)	10 (8.8%)	71 (8.1%)	13 (8.8%)	X0.13=2	0.94
	Panic Disorder (PD) , n(%)	22 (1.9%)	1	21	0	-	-
	Social Phobia, n(%)	38 (3.3%)	10 (8.8%)	26 (3.0%)	2 (1.4%)	X10.1=2*	0.005
	Agoraphobia without PD, n(%)	27 (2.4%)	6 (5.3%)	19 (2.2%)	2 (1.4%)	X5.01=2*	0.10
	Agoraphobia with PD, n(%)	5 (0.4%)	0	3	0	-	-
	Number of anxiety disorders, mean (SD)	0.16 (0.41)	0.24 (0.50)	0.16 (0.41)	0.12 (0.31)	F=2.95	0.053

SD=standard deviation; MI=Myocardial Infarction; LVEF=left ventricular ejection fraction.
* For X2 tests with cell-counts<5, Fisher's exact test is reported.
** Individual item-level CIDI symptoms were available for 1112 participants (98.0%)

Checking Data Quality

The non-parametric IRT-plots are shown in Supplement 1. The response behavior on most items did not show violations with respect to the expected form of the curves. However, for items 18 and 19, responses showed no meaningful relationship with underlying severity: higher categories did not become more likely to be endorsed as severity increased. Therefore, items 18 and 19 were removed from the subsequent analyses.

EFA was conducted next. The Eigenvalues of the first 4 factors were respectively: 7.84, 1.89, 1.02 and 0.90. The ratio of the first to second Eigenvalue was 4.14 and the ratios of the first to third and first to fourth Eigenvalues were even larger. These results indicated that the first factor by far explained the most common variance and, thus, that additional dimensions did not explain much additional variance. These results are in line with bifactor modeling results by Brouwer et al. [44], showing that the total score of the updated BDI (the BDI-II) adequately reflected overall depression severity and that only limited variance was explained by additional domain-specific factors.

TABLE 2. Item response theory item-parameters for the Beck Depression Inventory in a sample of patients with acute cardiac syndromes (n=1135) ordered by mean item-threshold-value.

BDI item	slope	Threshold 1*	Threshold 2*	Threshold 3*	Mean threshold
17-Fatigability	1.13	-1.97	1.62	3.35	1.00
15-Work inhibition	1.53	-0.78	1.30	2.57	1.03
16-Sleep problems	0.95	-0.51	1.59	2.84	1.31
21-Loss of libido	0.80	-0.10	1.74	2.88	1.51
13-Indecisveness	1.65	0.20	1.42	3.33	1.65
20-Somatic preoccupation	1.56	-0.14	1.83	3.59	1.76
2-Pessimism	2.52	1.17	1.88	2.43	1.83
4-Lack of satisfaction	2.08	-0.01	2.46	3.08	1.84
11-Irritability	1.06	0.32	2.54	2.84	1.90
1-Mood	1.89	0.78	2.20	3.90	2.29
6-Sense of punishment	1.09	1.99	2.49	2.52	2.33
10-Crying	1.12	0.89	3.06	3.15	2.37
3-Sense of failure	1.88	1.58	2.53	3.68	2.60
5-Guilty feelings	1.86	1.51	2.73	3.59	2.61
8-Self accusation	1.56	1.27	3.08	4.04	2.80
14-Body image	1.41	1.95	2.95	3.80	2.90
7-Self dislike	1.89	1.28	3.62	4.07	2.99
12-Social withdrawal	1.49	1.37	3.20	5.06	3.21
9-Suicidal thoughts	1.68	2.06	3.76	4.42	3.41

Parameters based on a graded response IRT model.
* The category thresholds represent the level of depression (theta) necessary to report the category or higher.

Item Response Theory

The IRT parameters are shown in Table 2. All items at the lowest end of the (theta) severity dimension (i.e. with the lowest mean thresholds) covered somatic symptoms and/or functional impairments. 'Fatigability' (item 17), 'work inhibition' (item 15), 'sleep problems' (item 16), 'loss of libido' (item 21), 'indecisiveness' (item 13),'somatic preoccupation' (item 20) and 'lack of satisfaction' (item 4) had mean thresholds ranging from 1.00 to 1.84. Items that covered depressive mood and/or depressive cognitions (e.g. 'depressed mood' [item 1], 'guilty feelings' [item 5], 'social withdrawal' [item 12] and 'suicidal thoughts' [item 9]) were all located higher on the severity dimension with average thresholds ranging from 2.29 to 3.42. Similar ordering of mean thresholds was seen when the IRT model was fitted in the DepreMI and MIND-IT subsamples separately (see Supplement 2). These results indicated that somatic/functional BDI items were reported at lower severity levels and that mood/cognitive BDI items were only reported at higher severity levels.

Person-fit

Person-fit distribution plots are shown in Figure 1. The frequency distribution of person-fit showed some negative skewness with a tail extending into the lower person-fit range. The distribution was quite uniform across increasing levels of severity (theta), although very low person-fit was more common at higher severity levels. Prior to regression analyses, person-fit was transformed as follows: -1*(ln($l_z$max minus l_z)). The results of the regression analyses with person-fit as dependent variable and the external variables as independent variables are shown in Table 3. Age and hypertension were positively associated and smoking was negatively associated with person-fit. In addition, person-fit was negatively associated with the number of reported CIDI depressive symptoms (beta=-0.111), the presence of several individual CIDI symptoms (e.g. 'feeling worthless/guilty' [beta=-0.110], 'depressive mood' [beta=-0.089] and 'psychomotor problems' [beta=-0.081]), CIDI social phobia, agoraphobia, and the total number of present anxiety disorders. In the multivariable analyses, only smoking, 'feeling worthless/guilty', 'depressive mood' and 'psychomotor problems' were negatively associated with person-fit. These results indicated that lower adherence to the group-based IRT model was associated with more features of clinical depression.

TABLE 3. Regression analyses to investigate the associations of externally determined variables with person-fit (outcome[1]) in a group of patients with an acute coronary syndrome (N=1135).

		Univariable[2]		Multivariable[3]	
		p	**beta**	**p**	**beta**
Demographic	Female gender	-0.01	0.71	-	-
	Age	0.11	<0.001	0.05	0.21
Cardiac severity	LVEF <45	0.00	0.97	-	-
	Killip class ≥2	-0.015	0.86	-	-
	Anterior site of MI	-0.04	0.20	-	-
Cardiac vulnerability	History of MI	0.02	0.59	-	-
	History of cerebral vascular disease	-0.03	0.32	-	-
	History of peripheral vascular disease	-0.01	0.88	-	-
	Family history of coronary artery disease	-0.04	0.14	-	-
Somatic health	Diabetes	-0.00	0.89	-	-
	Hypertension	0.06	0.04	0.04	0.28
	Hypercholesterolemia	-0.02	0.48	-	-
	Body mass index, mean (SD)	-0.01	0.80	-	-
	Current smoking	-0.12	<0.001	-0.08	0.03
CIDI depression after MI	Depression present after MI	-0.05	0.08	-0.03	0.60
	Number of CIDI depressive symptoms	-0.11	<0.01	-0.18	0.14
Individual symptoms after MI	Depressive mood	-0.09	<0.01	0.11	0.04
	Anhedonia/loss of interest	-0.03	0.32	-	-
	Energy loss	-0.05	0.11	-	-
	Appetite loss	-0.01	0.65	-	-
	Sleeping problems	-0.07	0.01	-0.05	0.33
	Psychomotor agitation/retardation	-0.08	0.01	-0.10	0.04
	Feeling Worthless/Guilty	-0.11	<0.001	-0.10	0.03
	Loss of Self-esteem	-0.05	0.08	0.02	0.70
	Concentration problems	-0.06	0.06	-0.06	0.22
	Preoccupations with death/suicide	-0.05	0.10	-	-
	#of previous depressive episodes	-0.05	0.08	-0.04	0.20
CIDI Anxiety after MI	Generalized Anxiety Disorder	0.02	0.51	-	-
	Panic Disorder (PD)	-0.06	0.06	-0.07	0.05
	Social Phobia	-0.07	0.03	-0.07	0.07
	Agoraphobia	-0.06	0.05	-0.06	0.11
	Agoraphobia with PD	-0.04	0.18	-	-
	# of anxiety disorders present	-0.06	0.03	0.06	0.24

MI, Myocardial Infarction; LVEF, left ventricular ejection fraction; PTCA, percutaneous transluminal coronary angioplasty; CABG, coronary artery bypass graft (surgery).

[1] Person-fit l_z was log-transformed and multiplied by -1 to make sure that positive coefficients reflect positive associations and negative coefficients reflect negative associations (outcome = -1 * (ln (l_z max – l_z)).

[2] Transformed outcome mean=0.34 (SD=0.55; range: -1.57-1.89)

[3] Conducted in a subsample of participants that missed none of the covariates (n=924; 81.3%); transformed outcome mean= 0.38 (SD=0.55; range=-1.57-1.78)

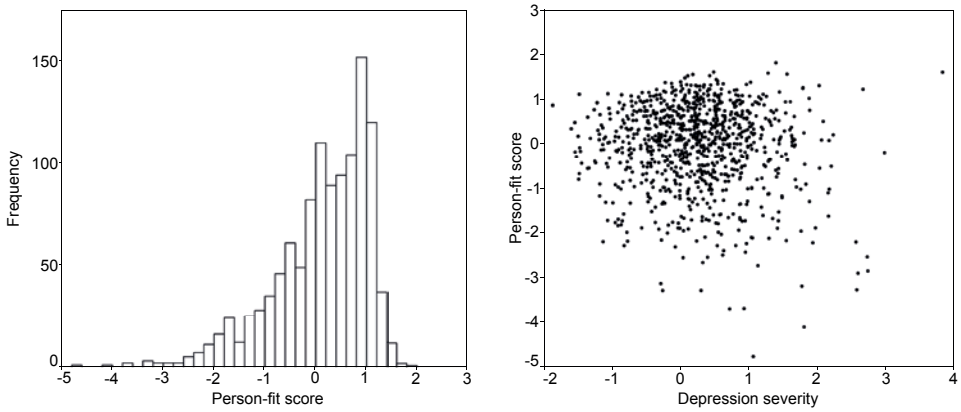


FIGURE 1. The person-fit frequency distribution (left) and the individual person-fit values (*y*-axis) plotted against severity (theta) according to the item response model (right).

Person-fit subgroups

The mean item-scores for the atypical and prototypical subgroups are shown in the upper panel of Figure 2. Most mean item scores were higher in the atypical group, in line with the finding that person-fit was inversely related with the number of depressive symptoms and the observation that the atypical group reported comparatively higher severity (theta). When these quantitative severity differences were adjusted by centering the mean scores in the atypical and prototypical groups on the middle group (Figure 2, lower panel), the groups' qualitative response pattern differences became more clearly visible. Compared to the prototypical group, the atypical group showed relatively higher scores on 'pessimism' (item2), 'sense of failure'(item 3), 'sense of punishment' (item 5), 'crying (item 10) and 'irritability' (item 11) and relatively lower scores on 'lack of satisfaction' (item 3), 'indecisiveness' (item 13), 'fatigability' (item17) and 'somatic preoccupation' (item 20). These results indicated that person-fit was related both to overall depression severity and specific patterns of item-endorsement.

Person-fit subgroups and external variables

The characteristics and comparisons of the subgroups are shown in Table 1. The atypical group was younger (mean age: 57.9 years) than the other groups, and the prototypical group was the oldest (mean age: 62.3 years). Of all health-related factors, only current smoking was more common in the atypical group. In addition, there were more patients with an actual CIDI MDD diagnosis in the atypical group and they more often reported 'feeling worthless/guilty' and 'loss of self-esteem'. Other CIDI symptom ratings did not differ across subgroups. Finally, there were more patients with a CIDI diagnosis of agoraphobia in the atypical group than in the other groups.

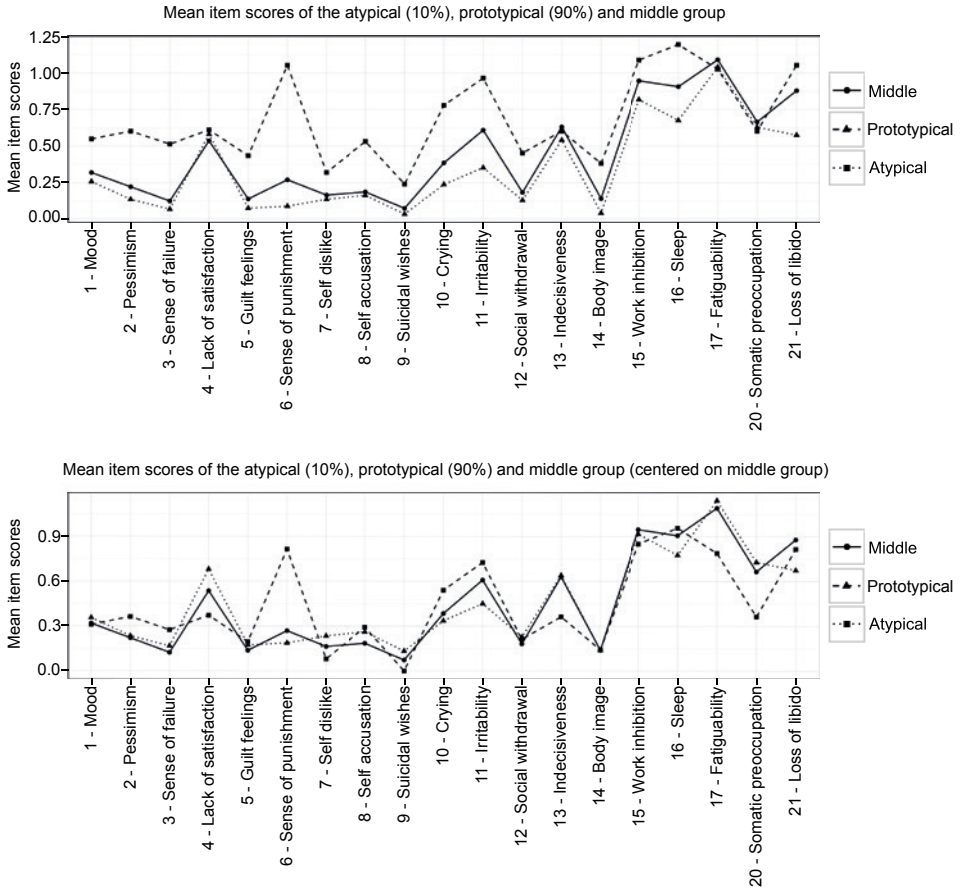


FIGURE 2. Mean item scores for the atypical, prototypical and middle person-fit groups. The upper panel displays the raw means, the lower panel displays the means after centering on the middle group (to eliminate quantitative severity differences and allow for better qualitative comparison of response patterns). Patients were allocated to the subgroups based on the 10% lowest (atypical) and 10% highest (prototypical) person-fit cutoff values. Because the observed person-fit was not normally distributed, a Monte Carlo simulation procedure was used to simulate an empirical person-fit distribution with the same sample characteristics as the observed data, from which unbiased person-fit cutoffs were derived.

DISCUSSION

This study was aimed to use a combination of group- and person-centered IRT approaches to investigate (1) whether in a sample of MI patients the BDI was biased towards measuring predominantly somatic/functional problems rather than actual depressive symptoms, and (2) whether it is possible to distinguish true depressive patients using person-fit. The group-based IRT results showed that somatic/functional items all had thresholds at the lower end of the severity spectrum, and were thus likely to be endorsed by most patients. Items covering depressive mood/cognitions all had higher thresholds and were thus less likely to be endorsed by most patients. Thus, at lower severity levels the BDI predominantly measures somatic symptoms, whereas at higher levels it measures depressive mood and cognitions. Variation in the adherence to this group-based IRT-model was investigated by calculating person-fit statistics. Low person-fit indicated that depressive mood/cognitions symptoms were endorsed, whereas somatic symptoms were not (or scarcely) endorsed (low adherence to the group-based IRT model) and regression analyses showed that this was positively associated with more CIDI depressive symptomatology and CIDI anxiety diagnoses. Comparison of person-fit subgroups (e.g. atypical vs. prototypical responders) showed that those with atypical response patterns were younger, smoked more often, were more likely to have a CIDI diagnosis of MDD or social phobia and reported more depressive symptoms on the CIDI. There were no differences in the rates of CIDI somatic symptoms. These results suggested that for most MI patients, BDI scores reflect the presence of somatic symptoms, whereas for atypical patients BDI scores more often reflect symptoms indicative of depression according to clinical classifications.

These results have several interesting implications. First, results indicate that the BDI predominantly measures somatic/functional symptoms in MI patients with low BDI scores, in line with previous suggestions[12–14]. Although the current study looked at the BDI, which is just one of several broadly used questionnaires, the identified measurement properties are unlikely to be unique to the BDI. Other broadly used questionnaires such as the Patient Health Questionnaire (PHQ-9 [45]) and BDI-II [46] also contain items that are likely to be over endorsed by somatic patients. However, an investigation of the BDI-II in MI patients showed that on this instrument, somatic symptoms were less over reported by MI patients, although, in line with the currently presented group-based IRT model, somatic symptoms did account for a majority (73.9%) of low (BDI-II <4) scores and this percentage decreased with increasing BDI-II scores (35.5% for those with BDI-II >12; [47]).

The findings suggest that in most MI patients, BDI scores reflect predominantly somatic problems. This is especially the case for those with relatively low BDI scores, since mainly somatic items were endorsed in the low severity range of the scale, in contrast to the mood/cognitive items. This relation between response-behavior and severity is important because previous studies on the association between depression and MI/ACS have often included 'depressed' patients based on relatively low scale cut-offs (e.g. BDI≥10 for mild-moderate depression[4,48]), which are lower than most clinical cut-offs (e.g. BDI≥10 for mild and BDI≥19 for moderate depression[27]). The current results (Table 2) suggest that a patient could meet such a cut-off by reporting almost exclusively somatic symptoms. Also, when using the BDI (or other instrument) as a continuous determinant, the range of the BDI scores is likely to influence the extent to which scores are biased due to over-reporting, with more somatic bias in the low severity range.

However, to state that the BDI always measures predominantly somatic symptoms in MI patients (and other somatic patient populations) would be too simplistic given the second implication of the results. The person-fit analyses showed that, despite the group-level IRT model, not all individual patients reported strictly somatic symptomatology. Indeed, lower person-fit was associated with increased key-features of depression and more comorbid anxiety, more indicative of the presence of psychopathology according to clinical classifications. If all BDI scores would be dismissed as being reflective of patients' somatic symptomatology and useless to detect depressive severity, these patients would be overlooked. This would be a shame because in any population of somatic patients there can be truly depressed patients that will respond accordingly on a depression questionnaire. Person-fit could help to distinguish such relevant cases from the majority of non-depressed MI-patients. If developed further, this could be of scientific and clinical interest, as it could help to improve the currently suboptimal specificity of questionnaires to detect true depressive cases.

The above suggests that person-fit could be helpful in clinical practice to identify potentially relevant cases. Once a group-based IRT-model has been established in a norm-population, computation of a person-fit statistics for individual patients is straightforward. An individualized person-fit statistic can then be used to evaluate whether a patient responds in a way that is atypical for the population he/she belongs to and could indicate the need for closer scrutiny of the individual item responses. Especially when questionnaires are administered digitally, such a procedure could be implemented quite easily.

The IRT findings are of conceptual interest in the light of previously developed dimensional approaches to distinguish between the somatic and mood/cognitive aspects of depression in psychosomatic research[49]. Some studies have used factor analyses to extract distinct factors and have used these in psychosomatic research[50]. There is an ongoing discussion about whether factor models of depression should take the form of a factor model with co-existing (correlated) factors[50] or that one should use a hierarchical, bifactor approach[11]. The current IRT study did not look deeply into the matter of (multi)dimensionality, but did indicate that there may exist yet another distinction between somatic vs. mood/cognitive items, with one cluster of items (somatic) positioned along the lower range of the severity spectrum and a second cluster (mood/cognitive) along the upper range. This indicates that in MI patients, item-clustering is related to the severity of the reported BDI score. Presumably, in populations without somatic illnesses, item-thresholds could be more evenly distributed along the severity dimension, making such clustering effects less likely.
Interestingly, we found no differences between the prototypical and atypical subgroup on measures of cardiac vulnerability such as LVEF, Killip class, and history of MI. These findings appear to be in contrast with a previous study on these data that showed stronger associations between cardiac vulnerability measures and the somatic depressive symptoms dimension compared to the cognitive/affective symptom dimension[50]. On the other hand, another report on data from the MIND-IT study showed that a lower LVEF was not only associated to higher baseline BDI scores, but also to increased rates of a depression diagnosis in the year following MI[51], suggesting that a reduced LVEF is not only characteristic for patients exhibiting somatic/functional symptoms (as seen in the prototypical sub group), but also for those with a formal depression diagnosis (as seen in the atypical subgroup).

The present results could explain previous findings from studies on the association between depression and MI/ACS outcome. For instance, one study showed that when adjusting for BDI score, previously observed associations between clinical depression and poor cardiac outcomes diminished, whereas BDI score remained predictive of cardiac outcomes[52]. Another study among ACS patients found that an overall depression severity measure (HADS-D) was associated with mortality, whereas the BDI Fast-Screen, which includes only the mood/cognitive items of depression, was not. Again, the full severity instrument could reflect a large amount of somatic symptom severity. Both studies' findings could be explained by the current finding that the BDI score is a somatic severity indicator in most patients, and thus, a good predictor of poor cardiac outcome. In line with this, the observation that even minimally increased BDI scores are associated with increased mortality among MI-patients[6] could be explained by the current finding that low BDI scores reflect predominantly somatic symptomatology.

The current study had several strengths, including the modern psychometric analyses, large sample size, and the fact that the psychometric results could be linked to clinically relevant external factors. However, several study limitations should also be kept in mind when interpreting the results. First, patients with a BDI<10 in the MIND-IT sample were excluded because they had no CIDI assessment, potentially leading to selection bias. However, IRT-analyses in all patients with and without a CIDI assessment (n=2469) showed a similar ordering of low item thresholds for somatic items and higher thresholds for depressive mood/cognition items (results not shown), indicating that the effect of sample selection on the IRT result was limited. Second, the number of assessed external variables is limited and other relevant determinants of person-fit could exist. Third, the results apply to the BDI and, although similar effects would be expected, the generalizability of the results to other, more recently developed instruments (e.g. BDI-II [46]; Inventory of Depressive Symptomatology [53]) should be investigated. Future research could focus on such replication efforts but also on further model development to better differentiate between groups of patients based on their response tendencies. Also, to find out whether the clustering of somatic symptoms at the lower end of the severity depression spectrum purely reflects somatic disease or also the presence of a latent depression subtype, future research should repeat the current analyses in a somatically healthy sample. The current (atypical/prototypical) subgroups were shown to be different in terms of psychopathology characteristics (e.g. depression diagnoses), but differentiation between distinct types of patients based on their item reporting could be further improved by use of a data-driven mixture approach to person-fit or mixture IRT[54]. Eventually, such an approach could help to even better distinguish those that report increased depression scores as an expression of mental health problems from the patients that report increased depression scores as an expression of their somatic illness.

In conclusion, the results indicate that the BDI measures predominantly somatic symptoms in the majority of MI patients with low BDI scores. However, person-fit could be used to identify a subgroup of patients with atypical response patterns and for whom BDI scores were indicative of clinical depression, which showed the potential usefulness of person-fit statistics for clinical purposes. In addition, the overall findings illustrate the potential usefulness of a person-centered IRT approach to gain more insight in symptom-specific link between depression measurements and cardiac outcomes.

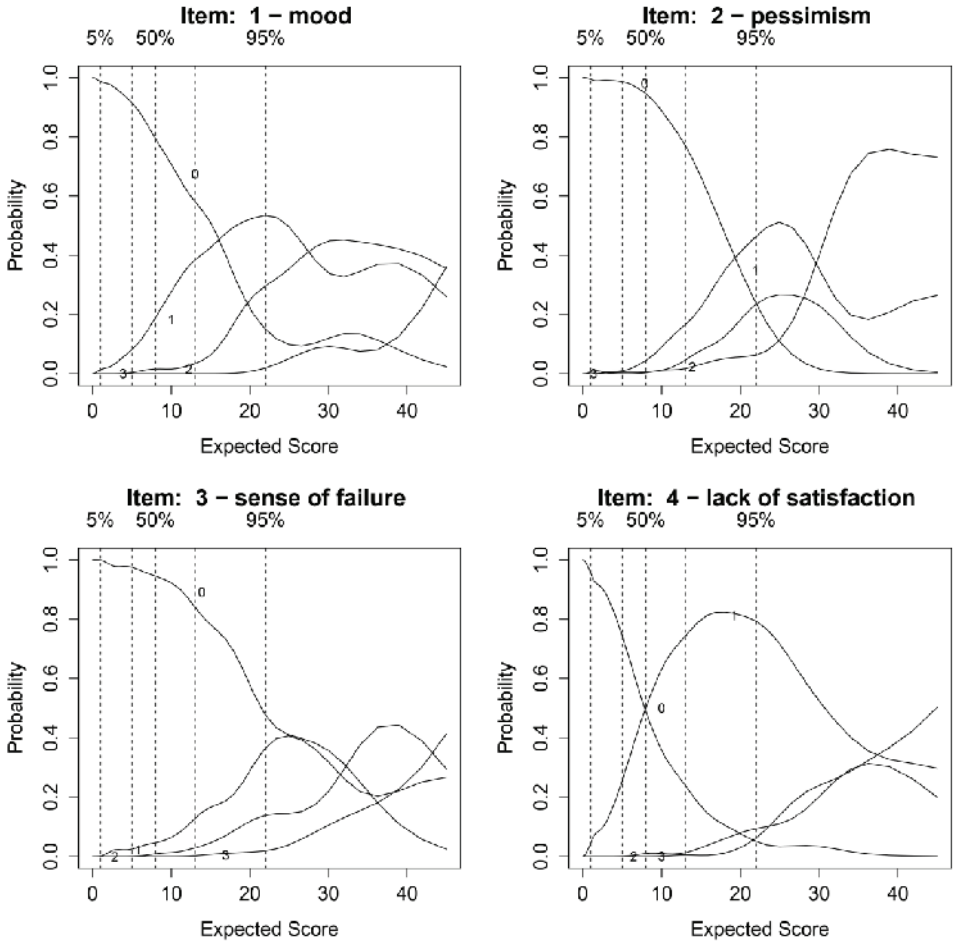
REFERENCES

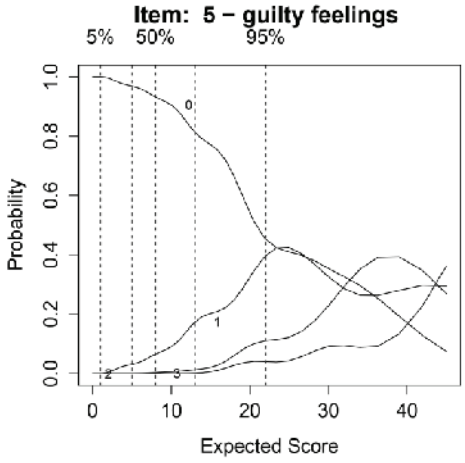
1. Van der Kooy, K. *et al.* Depression and the risk for cardiovascular diseases: systematic review and meta analysis. *Int. J. Geriatr. Psychiatry* **22,** 613–626 (2007).
2. Meijer, A. *et al.* Adjusted prognostic association of depression following myocardial infarction with mortality and cardiovascular events: individual patient data meta-analysis. *Br. J. Psychiatry J. Ment. Sci.* **203,** 90–102 (2013).
3. Frasure-Smith, N., Lespérance, F. & Talajic, M. Depression Following Myocardial Infarction: Impact on 6-Month Survival. *JAMA* **270,** 1819–1825 (1993).
4. Frasure-Smith, N., Lespérance, F. & Talajic, M. Depression and 18-Month Prognosis After Myocardial Infarction. *Circulation* **91,** 999–1005 (1995).
5. Carney, R. M. *et al.* Depression as a risk factor for mortality after acute myocardial infarction. *Am. J. Cardiol.* **92,** 1277–1281 (2003).
6. Bush, D. E. *et al.* Even minimal symptoms of depression increase mortality risk after acute myocardial infarction. *Am. J. Cardiol.* **88,** 337–341 (2001).
7. Thombs, B. D. *et al.* Prevalence of Depression in Survivors of Acute Myocardial Infarction. *J. Gen. Intern. Med.* **21,** 30–38 (2006).
8. Martens, E. J., Smith, O. R. F., Winter, J., Denollet, J. & Pedersen, S. S. Cardiac history, prior depression and personality predict course of depressive symptoms after myocardial infarction. *Psychol. Med.* **38,** 257–264 (2008).
9. Koenig, H. G., George, L. K., Peterson, B. L. & Pieper, C. F. Depression in medically ill hospitalized older adults: prevalence, characteristics, and course of symptoms according to six diagnostic schemes. *Am. J. Psychiatry* **154,** 1376–1383 (1997).
10. Sørensenf, C., Friis-Hasché, E., Haghfelt, T. & Bech, P. Postmyocardial infarction mortality in relation to depression: a systematic critical review. *Psychother. Psychosom.* **74,** 69–80 (2005).
11. Thombs, B. D., Ziegelstein, R. C., Beck, C. A. & Pilote, L. A general factor model for the Beck Depression Inventory-II: Validation in a sample of patients hospitalized with acute myocardial infarction. *J. Psychosom. Res.* **65,** 115–121 (2008).
12. Delisle, V. C. *et al.* The influence of somatic symptoms on Beck Depression Inventory scores in hospitalized postmyocardial infarction patients. *Can. J. Psychiatry Rev. Can. Psychiatr.* **57,** 752–758 (2012).
13. Leentjens, A. F. G., Verhey, F. R. J., Luijckx, G.-J. & Troost, J. The validity of the Beck Depression Inventory as a screening and diagnostic instrument for depression in patients with Parkinson's disease. *Mov. Disord.* **15,** 1221–1224 (2000).
14. Moran, P. J. & Mohr, D. C. The Validity of Beck Depression Inventory and Hamilton Rating Scale for Depression Items in the Assessment of Depression Among Patients with Multiple Sclerosis. *J. Behav. Med.* **28,** 35–41 (2005).
15. Chilcot, J. *et al.* A confirmatory factor analysis of the beck depression inventory-II in end-stage renal disease patients. *J. Psychosom. Res.* **71,** 148–153 (2011).
16. Embretson, S. & Reise, S. *Item Response Theory for Psychologists*. (Psychology Press, 2000).
17. Wanders, R. B. K. *et al.* Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make? *J. Psychosom. Res.* **78,** 130–136 (2015).
18. Meijer, R. R. & Sijtsma, K. Methodology Review: Evaluating Person Fit. *Appl. Psychol. Meas.* **25,** 107–135 (2001).
19. Meijer, R. R. Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychol. Methods* **8,** 72–87 (2003).
20. Zuidersma, M., Ormel, J., Conradi, H. J. & Jonge, P. de. An increase in depressive symptoms after myocardial infarction predicts new cardiac events irrespective of depressive symptoms before myocardial infarction. *Psychol. Med.* **42,** 683–693 (2012).
21. Bot, M., Pouwer, F., Zuidersma, M., van Melle, J. P. & Jonge, P. de. Association of Coexisting Diabetes and Depression With Mortality After Myocardial Infarction. *Diabetes Care* **35,** 503–509 (2012).

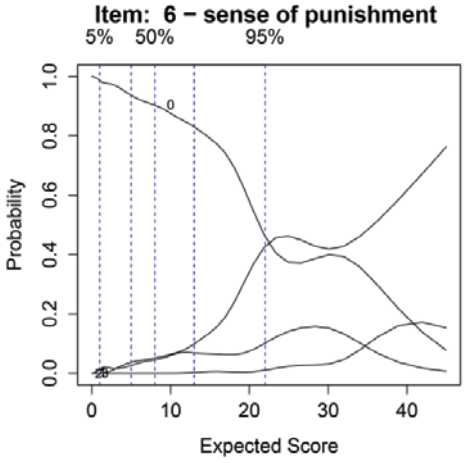
22. van den Brink, R. H. S. *et al.* Treatment of depression after myocardial infarction and the effects on cardiac prognosis and quality of life: Rationale and outline of the Myocardial INfarction and Depression-Intervention Trial (MIND-IT). *Am. Heart J.* **144,** 219–225 (2002).
23. van Melle, J. P. *et al.* Effects of antidepressant treatment following myocardial infarction. *Br. J. Psychiatry* **190,** 460–466 (2007).
24. Spijkerman, T. A. *et al.* Decreased impact of post-myocardial infarction depression on cardiac prognosis? *J. Psychosom. Res.* **61,** 493–499 (2006).
25. World Health Organization. *Composite International Diagnostic Interview (CIDI)*. (WHO, 1990).
26. Beck, A. T., Ward, C. & Mendelson, M. Beck Depression Inventory (BDI). *Arch. Gen. Psychiatry* 561–571 (1961).
27. Beck, A. T., Steer, R. A. & Carbin, M. G. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clin. Psychol. Rev.* **8,** 77–100 (1988).
28. World Health Organization. *Composite International Diagnostic Interview (CIDI, version 2.1. World Health Organization)*. (WHO, 1997).
29. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. (WHO, 1993).
30. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17,** 520–525 (2001).
31. Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. *impute: Imputation for microarray data.* (2013).
32. Meijer, R., Tendeiro, J. & Wanders, R. in *Handbook of item response theory modeling: Applications to typical performance assessment* 85–110 (Routledge, 2014).
33. Mazza, A., Punzo, A. & McGuire, B. *KernSmoothIRT: Nonparametric item response theory*. (2013).
34. Ramsay, J. O. Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika* **56,** 611–630 (1991).
35. Meijer, R. R. & Baneke, J. J. Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychol. Methods* **9,** 354–368 (2004).
36. Sijtsma, K. & Molenaar, I. *Introduction to nonparametric item response theory*. (SAGE, 2002).
37. Muthén, L. K. & Muthén, B. O. *Mplus User's Guide. 5th ed.* (Muhtén & Muthén, 1998).
38. Reise, S., Scheines, R., Widaman, K. & Haviland, M. G. Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling A Bifactor Perspective. *Educ. Psychol. Meas.* 5–26 (2013).
39. Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychom. Monogr. Suppl.* **34,** 100 (1969).
40. Rizopoulos, D. ltm: An R package for latent variable modelling and item response theory analyses. *J. Stat. Softw.* 1–25 (2006).
41. Reise, S. P. & Waller, N. G. How many IRT parameters does it take to model psychopathology items? *Psychol. Methods* **8,** 164–184 (2003).
42. Drasgow, F., Levine, M. V. & Zickar, M. J. Optimal Identification of Mismeasured Individuals. *Appl. Meas. Educ.* **9,** 47–64 (1996).
43. Seo, D. & Weiss, D. lz Person-Fit Index to Identify Misfit Students with Achievement Test Data. *Educ. Psychol. Meas.* **73,** 994–1016 (2013).
44. Brouwer, D., Meijer, R. R. & Zevalkink, J. On the factor structure of the Beck Depression Inventory–II: G is the key. *Psychol. Assess.* **25,** 136–145 (2013).
45. Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: Validity of a Brief Depression Severity Measure. *J. Gen. Intern. Med.* 606–613 (2001).
46. Beck, A. T., Steer, R. A. & Brown, G. K. *BDI-II manual*. (The Psychological Corporation, 1996).
47. Thombs, B. D. *et al.* Somatic symptom overlap in Beck Depression Inventory–II scores following myocardial infarction. *Br. J. Psychiatry* **197,** 61–66 (2010).
48. Frasure-Smith, N., Lespérance, F., Juneau, M., Talajic, M. & Bourassa, M. G. Gender, depression, and one-year prognosis after myocardial infarction. *Psychosom. Med.* **61,** 26–37 (1999).
49. Ormel, J. & de Jonge, P. Unipolar depression and the progression of coronary artery disease: toward an integrative model. *Psychother. Psychosom.* **80,** 264–274 (2011).

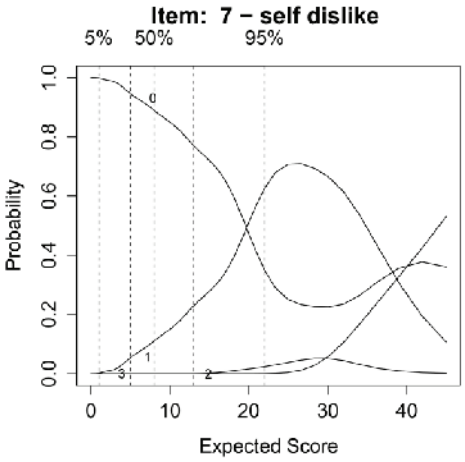
50. de Jonge, P. *et al.* Symptom Dimensions of Depression Following Myocardial Infarction and Their Relationship With Somatic Health Status and Cardiovascular Prognosis. *Am. J. Psychiatry* **163,** 138–144 (2006).
51. van Melle, J. P. *et al.* Relationship between left ventricular dysfunction and depression following myocardial infarction: data from the MIND-IT. *Eur. Heart J.* **26,** 2650–2656 (2005).
52. Zuidersma, M., Conradi, H. J., van Melle, J. P., Ormel, J. & de Jonge, P. Self-reported depressive symptoms, diagnosed clinical depression and cardiac morbidity and mortality after myocardial infarction. *Int. J. Cardiol.* **167,** 2775–2780 (2013).
53. Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B. & Trivedi, M. H. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol. Med.* **26,** 477–486 (1996).
54. Cohen, A. S. & Bolt, D. M. A Mixture Model Analysis of Differential Item Functioning. *J. Educ. Meas.* **42,** 133–148 (2005).

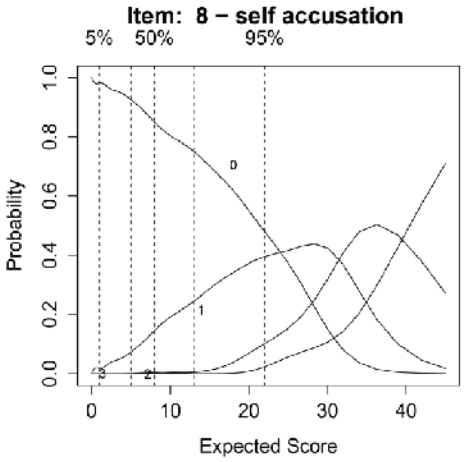
SUPPLEMENT 1. Non-parametric Item Response Analysis of the BDI items in a group of patients with an Acute Coronary Syndrome (n=1135). Category response functions (CRF) are plotted for each item using the KernSmoothIRT package. A CRF gives the probability of endorsing each response category given the expected total score. For good functioning items, the curves should have steep trace lines with narrow peaks, and the order of categories should correctly reflect severity at each interval of the total score. That is, at increased depression severity it should be more likely to endorse higher categories. The vertical lines represent the scores below which 5%, 25%, 50%, 75%, and 95% of patients fall. Lack of information can cause the CRF to be less well defined at the right side of the 95% vertical line.

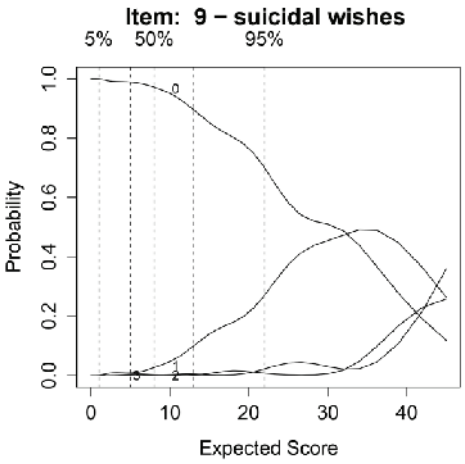


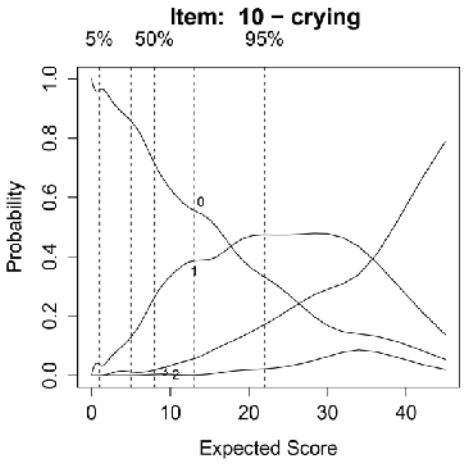

Item: 5 – guilty feelings
5% 50% 95%
Probability
Expected Score

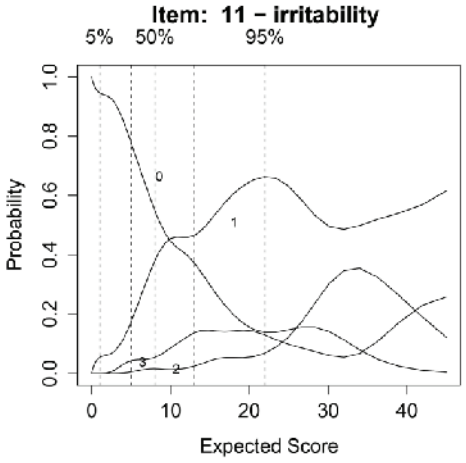

Item: 6 – sense of punishment
5% 50% 95%
Probability
Expected Score

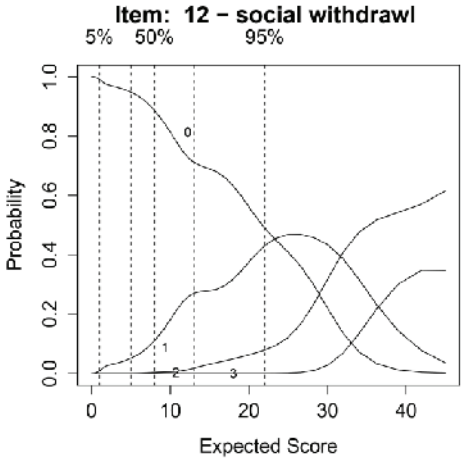

Item: 7 – self dislike
5% 50% 95%
Probability
Expected Score

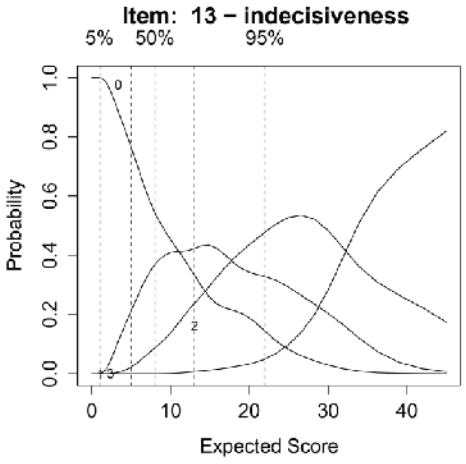

Item: 8 – self accusation
5% 50% 95%
Probability
Expected Score

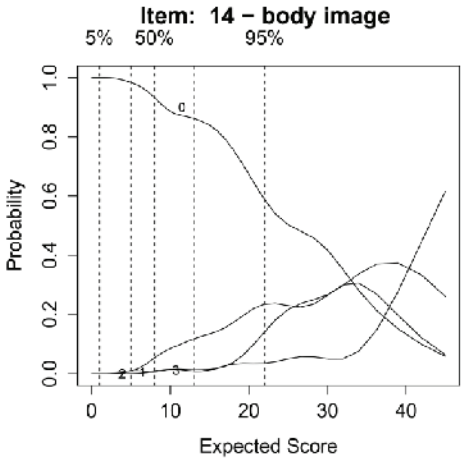

Item: 9 – suicidal wishes
5% 50% 95%
Probability
Expected Score

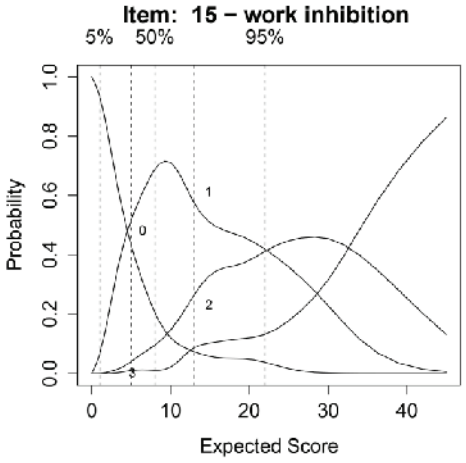

Item: 10 – crying
5% 50% 95%
Probability
Expected Score

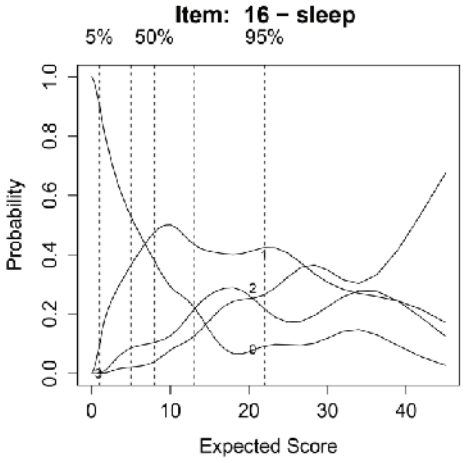

Item: 11 – irritability
5% 50% 95%
Probability
Expected Score

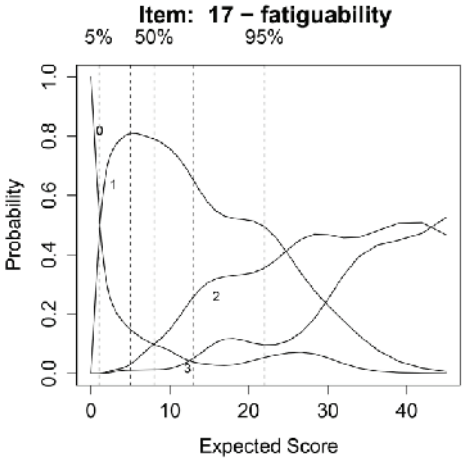

Item: 12 – social withdrawl
5% 50% 95%
Probability
Expected Score

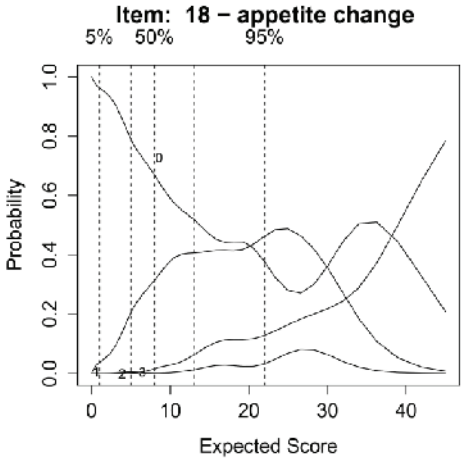

Item: 13 – indecisiveness
5% 50% 95%
Probability
Expected Score

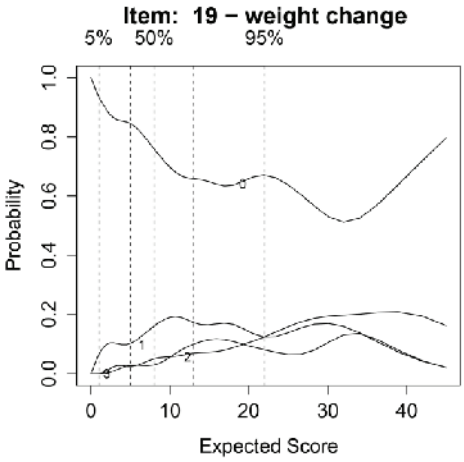

Item: 14 – body image
5% 50% 95%
Probability
Expected Score

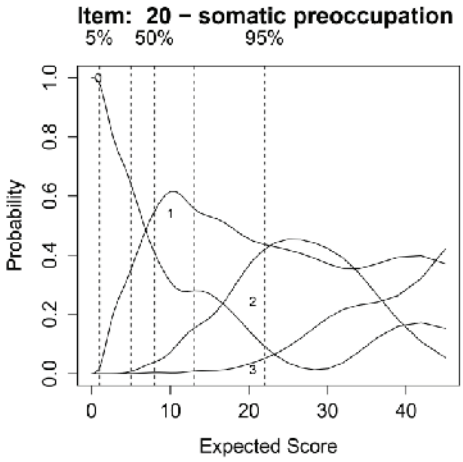

Item: 15 – work inhibition
5% 50% 95%
Probability
Expected Score

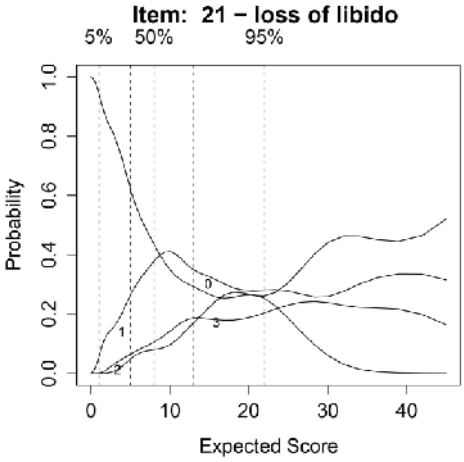

Item: 16 – sleep
5% 50% 95%
Probability
Expected Score


Item: 17 – fatiguability
5% 50% 95%
Probability
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3
0 10 20 30 40
Expected Score


Item: 18 – appetite change
5% 50% 95%
Probability
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3
0 10 20 30 40
Expected Score


Item: 19 – weight change
5% 50% 95%
Probability
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3
0 10 20 30 40
Expected Score


Item: 20 – somatic preoccupation
5% 50% 95%
Probability
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3
0 10 20 30 40
Expected Score


Item: 21 – loss of libido
5% 50% 95%
Probability
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3
0 10 20 30 40
Expected Score

SUPPLEMENT 2. When fitting a Graded Response Model (GRM) to the combined BDI data of Myocardial Infarction patients in the DepreMI and MIND-IT studies, the 'somatic/functional impairment' items were observed to cluster at the lower end of the severity dimension (theta), whereas items covering depressive cognitions were all located at the severe end of the dimension. To investigate the consistency of this clustering effect, the GRM model was fit separately in the DepreMI (n=424) and MIND-IT (n=711) samples and the order of the mean category thresholds was compared between the samples. The results are provided in the table below. Comparison of the ordering of the meanthresholds indicated that the observation of somatic/functional impairment items at the lowest end of the severity dimension was consistent across the independent study samples.

BDI items ordered by their mean category threshold in the total sample (n=1135), the DepreMI subsample (n=424) and the MIND-IT subsample(n=711)

Total sample		**DepreMI**		**MIND-IT**	
Item	MT	Item	MT	Item	MT
17-Fatigability	1.00	**17-Fatiguability**	1.18	**17-Fatiguability**	0.26
15-Work inhibition	1.03	**15-Work inhibition**	1.26	**16-Sleep**	0.68
16-Sleep problems	1.31	**16-Sleep**	1.47	**15-Work inhibition**	0.68
21-Loss of libido	1.51	**21-Loss of libido**	1.56	**21-Loss of libido**	0.81
13-Indecisveness	1.65	**13-Indecisiveness**	2.01	**20-Somatic preoccupation**	1.08
20-Somatic preoccupation	1.76	2-Pessimism	2.02	**13-Indecisiveness**	1.09
2-Pessimism	1.83	**4-Lack of satisfaction**	2.10	11-Irritability	1.31
4-Lack of satisfaction	1.84	11-Irritability	2.17	**4-Lack of satisfaction**	1.51
11-Irritability	1.90	**20-Somatic preoccupation**	2.19	1-Mood	1.62
1-Mood	2.29	12-Social withdrawal*	2.32	10-Crying	1.69
6-Sense of punishment	2.33	1-Mood	2.60	6-Sense of punishment	1.86
10-Crying	2.37	3-Sense of failure	2.76	2-Pessimism	1.86
3-Sense of failure	2.60	8-Self accusation	3.13	8-Self accusation	1.97
5-Guilty feelings	2.61	7-Self-dislike*	3.34	3-Sense of failure	2.00
8-Self accusation	2.80	9-Suicidal wishes	3.35	12-Social withdrawal	2.02
14-Body image	2.90	6-Sense of punishment	3.47	5-Guilty feelings	2.04
7-Self dislike	2.99	10-Crying	3.63	14-Body image	2.10
12-Social withdrawal	3.21	5-Guilty feelings	3.82	7-Self-dislike	2.26
9-Suicidal thoughts	3.41	14-Body image	4.32	9-Suicidal wishes	2.50

MT= mean threshold; Parameters based on a graded response IRT model. Somatic items are printed in bold.
*) Not enough data to reliably estimate this threshold

CHAPTER 4

Differential Reporting of Depressive Symptoms across Distinct Clinical Subpopulations: What DIFference does it make?

Rob B. K. Wanders, Klaas J. Wardenaar,
Ronald C. Kessler, Brenda W.J.H. Penninx,
Rob R. Meijer, Peter de Jonge

Journal of psychosomatic research 2015,
78:130-136.

ABSTRACT

Objective. To investigate the impact of differences in depressive symptom reporting across clinical groups (healthcare setting, chronic illness, depression diagnosis and anxiety diagnosis) on clinical interpretability and comparability of depression scores.

Methods. Participants from the Netherlands Study of Depression and Anxiety (n=2981) completed the self-report Inventory of Depressive Symptomatology (IDS-SR). Differences in depressive symptom reporting between distinct clinical subpopulations were assessed using a Differential Item Functioning (DIF) analysis. The effects of DIF on symptom level were evaluated by examining whether DIF-adjustment had clinically relevant effects.

Results. Significant DIF was detected across all tested clinical subpopulation groupings. Clinically relevant DIF was found on the symptom level for 13 IDS-SR items. However, impact of DIF on the aggregate level ranged from small to negligible: adjustment for DIF only led to salient changes in aggregate scores for 0.2-12.7% of individuals across tested sources of DIF.

Conclusion. Differences in endorsement patterns of depressive symptoms were observed across clinical populations, challenging the assumptions regarding the measurement properties of self-reported depression. However, effects of DIF on the aggregate level of IDS-SR total scores were found to be minimal and not clinically important. The IDS-SR thus seems robust against DIF across clinical populations.

INTRODUCTION

Accurate assessment of the severity of depressive symptoms is important and requires items to have the same meaning for all persons[1]. If depression measures do not have equivalent meaning across subgroups, biased results may be obtained leading to interpretation difficulties. It is often implicitly assumed that self-report measures are 'invariant' and that valid comparisons can be made between persons or groups. However, psychometric work suggests that depression measures are often not invariant and symptoms are differently reported depending on the target population[2].

To date, differences in the reporting of depressive symptoms have been observed between patient populations with different clinical characteristics, such as different healthcare setting[3], and the presence of chronic somatic conditions or psychiatric comorbidities[4]. In such cases different depression scores do not merely indicate differences in depression because the measurement of depression may be confounded by clinical factors. Indeed, in an overview, Teresi et al.[5] showed that most studies report items to function differently across populations with sizeable magnitude and impact on depression measurement. This phenomenon is called differential item functioning (DIF), where patients from distinct groups with equal depression severity do not have the same probability of endorsing a given item.

Most studies[5] look at DIF between classical socio-demographic groups such as age, gender, and race. Although it is important to know the presence and impact of DIF across these groups, they may not be the most profound sources of DIF in depression measures. First, it has been argued that demographic groupings are quite arbitrary with regard to how accurately they identify persons who actually respond differentially to an item, and that theoretically motivated groupings would be of much greater interest[6]. Second, it is of particular importance to investigate DIF across clinically defined groups (e.g. patients vs. controls; primary vs. secondary care), as these are the subgroupings where comparisons of depression scales have pronounced scientific and clinical relevance.

We hypothesize two ways in which clinical characteristics may affect symptom reporting across clinical populations. First, increased prevalence of a symptom may be the result of being more common in one clinical population than in another population, without indicating differences in depressive severity. For example, patients in primary care have more comorbid medical illness and less comorbid mental disorders than patients treated in secondary care[3]. As a result, patients may have different probabilities of reporting particular symptoms of depression as they are confounded by the presence of symptoms of other disorders or complaints. Second, differences between patients in distinct clinical populations may be explained by response behavior. The social-cognitive

theory of survey response states that the process of responding to a questionnaire involves a step of organizing and articulating symptoms[7]. In this process, people interpret questions relative to others in the same setting and express symptoms in a way they feel is appropriate for them. Different clinical settings may form different frames of reference for the interpretation and expression of symptoms, resulting in qualitative differences in symptom reporting.

In the current study, item response theory (IRT) methods were applied to perform DIF analyses[8]. The benefit of this method is that it accounts for a potential confounding effect of depression severity in evaluating population differences. So far, most work has mainly focused on detection of DIF rather than on its actual practical impact[9]. The few studies that have investigated practical impact show mixed results[10]. Some found low practical impact[11,12], whereas others found substantial impact[13,14]. The aim of the current study was to investigate DIF across distinct clinical populations, and more importantly, to evaluate the extent to which this leads to clinically important differences. Analyses will be conducted on data collected with the Inventory of Depressive Symptomatology Self Report (IDS-SR) in a large cohort study.

METHODS

PARTICIPANTS AND PROCEDURES

Data were obtained from the Netherlands Study of Depression and Anxiety (NESDA), a large scale longitudinal cohort study among 2981 adult participants (1002 men and 1979 women) aged 18-65 at baseline assessment (2004-2007). The sample consisted of 2329 persons with a lifetime diagnosis of depressive or anxiety disorder and 652 persons without a lifetime psychiatric diagnosis. A detailed account of the rationale, objectives, and methods of NESDA can be found elsewhere[15]. All participants had a face-to-face structured interview with a trained research assistant, consisting amongst others of a standardized psychiatric and demographic interview, biomedical measurements, a blood-draw and a battery of self-report questionnaires. The protocol of NESDA was approved by the Ethical Committees of all participating universities and all subjects signed informed consent.

INSTRUMENTS

The IDS-SR[16] consists of 30 items, each with four response options (scored 0-3). As only ‘appetite increase’ or ‘appetite decrease’, and only ‘weight gain’ or ‘weight loss’ are scored, these items were combined into single ‘appetite change’ and ‘weight change’ items.

The Composite International Diagnostic Interview (CIDI, WHO version 2.1) was used to assess the presence of DSM-IV major depressive disorder, dysthymia, social phobia, generalized anxiety disorder, panic disorder and agoraphobia. The presence of chronic illness was assessed during the face-to-face structured interview.

STATISTICAL ANALYSES

Missing data

Only few participants had missing values on the IDS-SR. Forty participants who missed more than five responses were removed from the sample. For the remaining participants, missing values were imputed (399 item scores, 0.54%) using the R package 'impute' with default settings[17]. The imputation procedure uses a K-nearest neighbor (KNN) search to impute based on scores of subjects with similar symptom profiles as the subject with missing values. This method was chosen on theoretical considerations as differences in symptom reporting across clinically defined subpopulations were hypothesized and imputation based on the whole sample would contradict this.

Exploratory analysis

To inspect data quality and IRT assumptions[18] we performed nonparametric IRT analyses, which have been shown to provide excellent tools to explore data and get insight in the suitability of the data for parametric modeling[19,20]. Item response behavior was visually inspected using Testgraf[21] and IRT assumptions were inspected with MSP5.0[22] and the R package 'mokken'[23]. The strongest assumption underlying DIF analyses is unidimensionality, meaning that a single latent trait dimension (level of depression) underlies the probability of reporting a symptom. As empirical data will never be strictly unidimensional[24], the extent of unidimensionality was checked by fitting a bi-factor model with each item loading on a general factor and on a specific group factor. Group factors were specified using results from previous factor analytic studies on the IDS-SR[16,25]. Model fit was assessed with the root mean square error of approximation (RMSEA) and the comparative fit index (CFI) with RMSEA <0.06 and a CFI >0.95 indicating good fit[26]. The data is considered sufficiently unidimensional when the common factor is large enough such that the presence of smaller factors does not influence the estimation of the model, and thus results are not biased by the existence of multidimensionality[24]. Two criteria were used to assess whether the data was sufficiently unidimensional: (a) Factor loadings were compared between the general factor and the group factor[27]. (b) The explained common variance (ECV) was inspected[24]. This analysis was performed using the R package 'lavaan'[28] with adjusted weighted least squares (WLSMV) estimation.

Differential item functioning

An item shows DIF when subjects of distinct groups, with equal levels of depression, do not have the same probability of reporting a symptom (endorsing an item). There are two types of DIF. With uniform DIF, the strength of the effect remains constant across different levels of depression: one group consistently has a higher probability of reporting a symptom at equal levels of depression. With non-uniform DIF, the effect has a different strength or direction across different levels of depression; e.g. one group has a higher probability of endorsement at low levels of depression and a lower probability of endorsement at high levels of depression.

To detect DIF, a hybrid iterative technique was used that combines logistic regression and IRT. The R package 'lordif'[29] was used. Results were replicated with a different method (IRTPRO[30], results not shown). Logistic regression provides a flexible framework to detect both uniform and non-uniform DIF[31]. Three nested models are formed in hierarchy: the first model includes a direct effect for ability (level of depression), the second model includes a direct effect for group and the third model includes an interaction effect of group and ability. The presence of DIF is tested by comparing the log likelihood values of the first and second model (uniform DIF), the second and third model (non-uniform DIF) and the first and third model (overall DIF) using likelihood ratio x2 tests. Incorporating IRT in this framework allows for better estimation of the level of depression, to account for the effects of DIF, and to visually inspect the effects of DIF with group-specific item response functions.

The two parameter graded response model[32] was fit as an IRT model. The discrimination parameter (α) describes how strong a symptom (item) is related to the underlying depression severity (person characteristic), and the category thresholds (β) reflect the severity of a symptom (item). This model was appropriate for our purposes because item response data consisted of ordered categorical responses reflecting symptom severity[6]. For the items that were free of DIF, IRT parameters were obtained from the whole sample. These items functioned as anchor-items to place all individuals on the same scale, making the individual depression scores comparable. For items with DIF, group-specific IRT parameters were obtained and used to estimate individual scores that account for DIF. To control for false-positive or false-negative DIF findings, these results were used in subsequent purifying analyses until the same set of items was found to show DIF over successive iterations[29].

One problem of DIF analyses is the detection of statistically significant, yet clinically unimportant effects. Therefore, a-priori criteria were set to evaluate clinical relevance of detected DIF, using previous studies as guideline[9,10,29,31,33–35] (8,9,28,30,33,34,35). First, a Bonferroni correction was applied to correct for multiple testing across items (α=0.002).

Second, uniform DIF was only considered relevant if the threshold difference was at least 0.3. Third, McFadden pseudo R^2 impact measures were compared to empirical thresholds obtained from 10,000 simulated datasets under no-DIF conditions using Monte Carlo simulations in 'lordif'. The value corresponding to the 99.8% quantile that cuts the largest 0.2% (α=0.002) over all iterations was computed for each item. As thresholds may differ across items, the maximum value was taken as cutoff.

General outline of the analyses

DIF was analyzed with respect to: (1) healthcare setting (primary care vs. secondary care), (2) chronic illness (present vs. not present), (3) MDD diagnosis (present vs. not present), and (4) anxiety diagnosis (present vs. not present). Besides symptom-level DIF, the impact of DIF on the aggregate level was investigated by comparison of DIF-adjusted and unadjusted scores. A difference between adjusted and unadjusted scores of at least one standard error of measurement (SEM) was considered clinically important, as it closely approximates the minimal clinical important difference (MCID) [36,37]. To evaluate whether adjustment for DIF improved the diagnostic properties of the IDS-SR, receiver operating characteristic (ROC) analyses[38] were conducted with adjusted and unadjusted scores using CIDI MDD diagnosis as gold standard.

RESULTS

DESCRIPTIVES

Demographic and diagnostic information are shown in Table 1. Patients in the sample had a mean age of 41.9 (range 18-65), and 66.9% were female. Respondents treated in secondary care were younger ($F=89.36$, $p<0.001$), and they were more likely to have a current diagnosis of major depressive disorder ($x2=689.6$, $p<0.001$) and anxiety disorder ($x2=436.5$, $p<0.001$).

DATA QUALITY AND MODEL ASSUMPTIONS

Nonparametric IRT analyses revealed some items that violated model assumptions (full report available from RW). Inspection of item response functions showed that most violations were caused by rank-order problems, in which case a higher category was equally or more popular than a lower category. To improve item functioning, these items were recoded from four to three categories. After rescoring, the items 'onset insomnia', 'mid insomnia', 'morning insomnia', 'hypersomnia', and 'mood variation' still showed unsatisfactory response behavior and model violations and were excluded from further analyses.

TABLE 1. Demographic characteristics from the NESDA study (n=2981)

Demographic	n	%	Mean	SD
Gender				
Male	1002	33.6		
Female	1979	66.4		
Age	2981	100.0	41.9	13.1
Years of education	2981	100.0	12.2	3.3
Healthcare setting				
Primary care	1968	66.0		
Secondary care	1013	34.0		
CIDI diagnosis				
Current MDD	1115	37.4		
Lifetime MDD	1925	64.6		
Current anxiety	1305	43.7		
Lifetime anxiety	1772	59.4		
Chronic illness	1627	54.6		

SD, standard deviation.

The bi-factor model showed good fit (CFI=0.99;RMSEA=0.03). The average loading on the general factor was 0.37 (range=0.20-0.51). Three items had loadings below 0.3. Factor loadings on group factors were comparatively small, (average=0.1;range=-0.04-0.31). Only one item loaded above 0.3 and five items had negative loadings. The explained common variance (ECV) was 0.85, indicating that most covariance was explained by the general factor. These results were taken to indicate that the data were sufficiently unidimensional to warrant further DIF analyses.

SYMPTOM-LEVEL DIF

Items with relevant DIF are listed in Table 2. For these, symptom-level DIF was significant at α=0.002, threshold differences exceeded 0.3, and McFadden pseudo R^2 values were larger than the empirically obtained cutoff score. Relevant DIF was observed with respect to healthcare setting (9 items), chronic illness (4 items), MDD diagnosis (7 items), and anxiety diagnosis (8 items). In Table 2, the x2 values for the likelihood ratio tests of uniform and non-uniform DIF are shown. Three different effects could be distinguished and are illustrated in Figure 1. The item 'somatic complaint' showed Uniform DIF with the primary care group being more likely to report the symptom than the secondary care group at all levels of depression. The item 'self outlook' showed Non-uniform DIF with the secondary care group being more likely to report the symptom at low levels of depression and less likely at high levels of depression. In addition, the item 'reactivity of mood' showed a

combination of uniform and non-uniform DIF, where secondary care patients were more likely to endorse the item, but this effect became weaker at lower levels of depression.

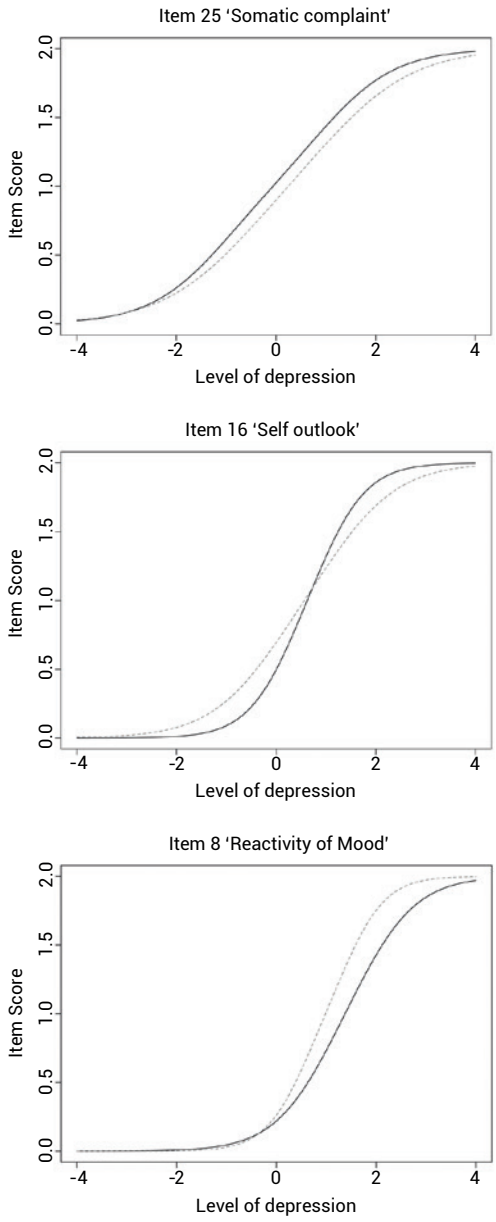


FIGURE 1. Expected item score functions showing the expected score on 'somatic complaint', 'self outlook', and 'reactivity of mood' for primary care (solid) and secondary care (dashed) at different levels of depression.

The impact of DIF on symptom level is illustrated with pseudo R^2 values (Figure 2), which reflect the impact of the observed DIF (uniform and non-uniform) on the distribution of individuals. Low values indicate that the effect of DIF occurs for levels of depression where only few individuals are located, whereas higher values indicate that the effect occurs at more common depression levels. Values are compared to empirical thresholds obtained from simulated datasets for healthcare setting (0.003), chronic illness (0.004), MDD diagnosis (0.004) and anxiety diagnosis (0.004).

TABLE 2. Item level uniform (U) and non-uniform (N) DIF of IDS-SR depressive symptoms for healthcare setting (primary care vs. secondary care), chronic illness (present vs. not present), MDD diagnosis (present vs. not present), and anxiety diagnosis (present vs. not present). x^2 values[1] are reported[2].

		Healthcare setting		Chronic illness		MDD diagnosis		Anxiety diagnosis	
		U	N	U	N	U	N	U	N
6	Irritable	3.1	24.0						
7	Anxious	15.7	23.2			16.2	12.0	154.2	13.1
8	Reactivity of mood	20.4	13.7			0	62.0	0	19.1
10	Quality of mood	8.4	37.4			8.4	130.4	7.1	79.0
16	Self outlook	0.9	42.5			1.6	81.6	38.6	36.8
23	Psychomotor slowing					1.6	44.8		
24	Psychomotor agitation	0.0	23.2			7.3	35.3	45.7	39.4
25	Somatic complaint	33.4	0.0	171.9	7.3				
26	Sympathetic arousal			41.0	0.3			69.0	1.2
27	Panic	22.3	15.8			30.7	10.7	285.7	15.7
28	Gastrointestinal			72.7	1.3				
29	Sensitivity	19.1	2.1					23.6	4.8
30	Leaden paralysis			25.5	0.1				

U, uniform DIF; N, non-uniform DIF.
[1] Differences between groups are evaluated using x2 test statistics with 1 degree of freedom.
[2] Cells were left empty if no relevant DIF was found.

Table 3 shows the threshold (β) and discrimination (α) item parameter differences across the groups for the items that were found to display relevant DIF. The first thresholds give the level of depression at which there is a 50% probability of reporting the symptom (reporting a score of 1 or higher). Low values indicate that the symptom is reported at low levels of depression. A negative difference between thresholds indicates that the first clinical group endorses the item at lower levels of depression than the second group, whereas a positive difference indicates vice versa. For instance, patients in primary care report 'somatic complaint' at lower levels of depression than patients in secondary care. The discrimination parameter describes the rate at which the probability of item endorsement

increases with increasing level of depression. Low values indicate only a small increase in the probability of reporting the symptom when the level of depression increases. A positive difference between discrimination parameters indicates that the symptom is more stable across levels of depression and less indicative for depression for the second clinical group compared to the first group. For example, patients with an anxiety diagnosis are more likely to endorse the 'self outlook' item at low levels of depression but less likely at high levels of depression compared to patients without an anxiety diagnosis. Consequently, the probability of reporting these symptoms remains constant as depression severity increases and is less characteristic for depression by anxiety patients. Negative differences would indicate vice versa.

Overall, clinically relevant DIF was found for 13 out of the 23 symptoms across all analyses. Combinations of uniform and non-uniform DIF were observed.

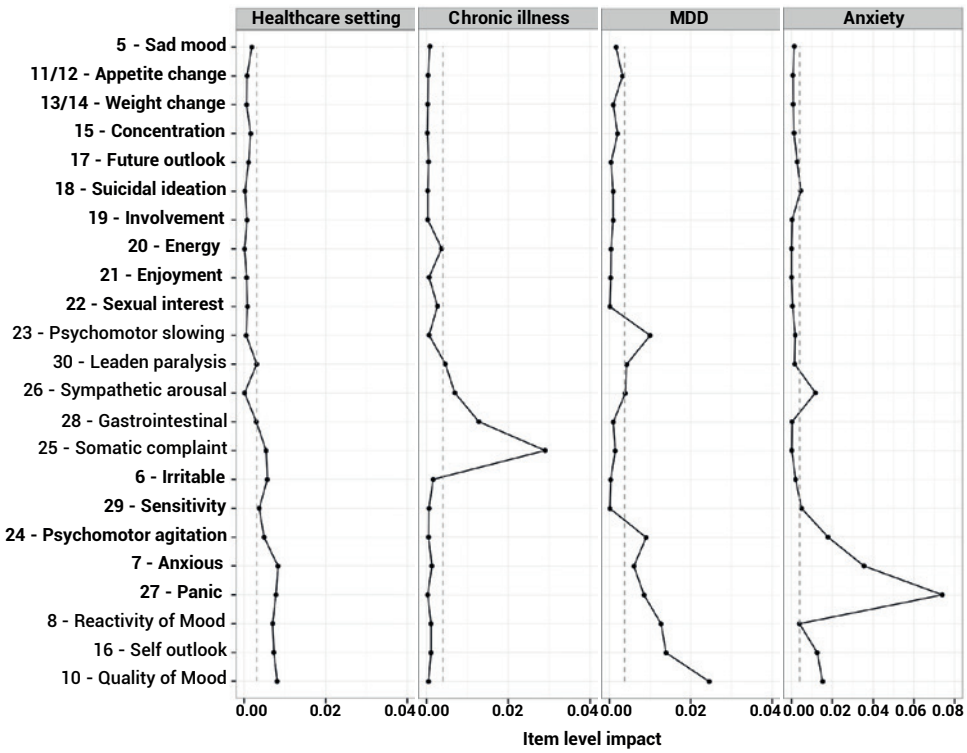


FIGURE 2. Plots of item level impact for each DIF analysis. McFadden pseudo R^2 values are plotted for each item (solid line) showing the magnitude and impact of DIF with empirical thresholds (dotted line). Items are clustered by item content and ordered by magnitude of DIF.

TABLE 3. Differential item functioning of IDS-SR depressive symptoms for healthcare setting (primary care vs. secondary care), chronic illness (present vs. not present), MDD diagnosis (present vs. not present), and anxiety diagnosis (present vs. not present). Differences in estimated first thresholds[3] and discrimination parameters between groups are reported[4].

		Healthcare setting		Chronic illness		MDD diagnosis		Anxiety diagnosis	
		β[1]	α[2]	β[1]	α[2]	β[1]	α[2]	β[1]	α[2]
6	Irritable	0.27	0.77						
7	Anxious	0.52	0.80			0.10	0.77	0.93	0.36
8	Reactivity of mood	0.40	-0.69			0.22	-2.58	0.17	0.19
10	Quality of mood	0.38	0.61			0.53	1.97	0.35	-0.41
16	Self outlook	0.37	0.65			0.35	1.09	0.47	0.80
23	Psychomotor slowing					-0.02	1.90		
24	Psychomotor agitation	0.20	0.55			0.05	1.15	0.69	0.68
25	Somatic complaint	-0.38	0.20	1.03	-0.04				
26	Sympathetic arousal			0.41	0.26			0.75	0.14
27	Panic	0.68	0.57			-0.15	0.63	3.10	0.39
28	Gastrointestinal			0.95	-0.06				
29	Sensitivity	0.46	0.30					0.48	0.40
30	Leaden paralysis			0.36	0.33				

β, threshold differences; α, discrimination parameter differences.

[1] First threshold differences resemble differences in probabilities of reporting a symptom. Positive differences indicate that the first clinical group endorses the item at lower levels of depression than the second group, whereas negative differences indicate vice versa.

[2] Positive differences in discrimination parameter indicate that the symptom is less indicative for depression for the second clinical group compared to the first group, whereas negative differences indicate vice versa.

[3] Second category threshold differences were found to be less informative and are not reported to increase interpretability and readability of the table.

[4] Cells were left empty if no relevant DIF was found.

AGGREGATE-LEVEL DIF

Impact of DIF on individual depression scores was assessed by taking the difference between the DIF-adjusted and unadjusted IRT estimated scores. Using this method, salient changes in individual depression scores (>1 SEM) after adjustment for healthcare setting-, MDD-, and anxiety-related DIF were only observed for a small number of patients. Adjustment for healthcare setting DIF led to salient change in only 6 patients (0.2%), for MDD DIF in 66 patients (2.2%), and for anxiety diagnosis DIF in 35 patients (1.2%). Adjustment for chronic illness related DIF led to a salient change in more participants 374 (12.7%).

Mean group-differences between adjusted scores were smaller than between unadjusted scores. Across primary and secondary care the mean difference was 0.95 between unadjusted and 0.90 between adjusted scores. This indicates that only 4.8% of

the observed difference in IDS-SR scores between healthcare settings is due to DIF. In a similar vein, 14.6% of IDS-SR differences between patients with and without a chronic illness was attributable to DIF, 8.1% between patients with and without MDD diagnosis, and 11.8% between patients with and without an anxiety diagnosis.

ROC analyses did not show differences between DIF-adjusted and unadjusted scores for any clinical group (all AUC=0.87) indicating no effect of DIF-adjustment on the diagnostic properties of the IDS-SR.

DISCUSSION

The aim of the present study was to investigate to what extent differences in depressive symptom reporting across clinical populations (healthcare setting, chronic illness, MDD diagnosis, anxiety diagnosis) led to clinically important differences in depression assessment.

Several preparatory analyses were run to inspect assumptions and data quality. Non-parametric analyses suggested the removal of several poor fitting items. The remaining items were rescored to three categories to improve item quality, in line with previous work on the same data[25]. Subsequent checks of unidimensionality showed data to be sufficiently unidimensional to perform DIF analyses.

SYMPTOM-LEVEL DIF

IDS-SR items showed meaningful DIF with respect to healthcare setting (9 items), chronic illness (4 items), MDD diagnosis (7 items), and anxiety diagnosis (8 items). Findings were consistent across analyses, with 13 out of 23 items showing DIF across one or more clinical groups. Interestingly, most DIF was observed on associated (non-criterion) depressive symptoms (e.g. 'reactivity of mood') and not on the DSM-criterion symptoms (e.g. 'sad mood'). Non-criterion symptoms of depression might be more related to confounding factors, and therefore more prone to DIF.

Alternatively, the current results could be explained by the fact that non-criterion symptoms may be more subjectively reported and support the hypothesis that different patients have different frames of reference, which may influence the interpretation and expression of symptoms. With respect to gender, self-reported depression responses have been shown to be heterogeneous in response scale usage and anchoring vignettes can narrow these gender differences[39]. Differences in interpretation and expression of symptoms between clinical groups could further be studied using a similar approach.

Somatic symptoms were found to show DIF across clinical groups. Participants with chronic illness were more likely to report the items 'somatic complaint', 'sympathetic

arousal', 'gastrointestinal', and 'leaden paralysis' than participants without chronic illness. Similar DIF was seen across healthcare settings. In primary care 'somatic complaint', and 'sympathetic arousal' were more likely to be reported than in secondary care. This could be caused by the tendency in primary care patients to report comparatively more somatic complaints than secondary care patients[3]. Alternatively, patients who mainly suffer from the somatic symptoms included in the IDS-SR may be more likely to end up in primary care than those who suffer from the full range of IDS-SR symptoms[40]. Patients with an anxiety diagnosis were more likely to report the 'sympathetic arousal' item than patients without anxiety diagnosis. Indeed, these patients have been found to have more medically unexplained symptoms than patients without anxiety disorder[41]. In addition, somatic hyperarousal has been shown to be a key-aspect of anxiety, especially of panic disorder[42].

Anxiety related symptoms showed DIF across clinical groups. Patients with an anxiety diagnosis were more likely to report the items 'anxious', 'psychomotor agitation', 'panic', and 'sensitivity' than those without anxiety. With respect to healthcare setting, 'irritable', 'anxious' and 'panic' were more likely to be reported in secondary care patients (vs. primary care) at low levels of depression than primary care patients. These items functioned more stably in secondary care (i.e. the probability of endorsement remained constant across the depression severity dimension) and, consequently, were better markers of depression in primary care. With respect to MDD diagnosis the items 'anxious', 'psychomotor slowing', 'psychomotor agitation', 'panic', and 'sensitivity' were more likely to be reported by patients with a diagnosis at low levels of depression than patients without a MDD diagnosis. Again, the items were more stable in diagnosed patients and more indicative for depression in undiagnosed patients. These findings could result from the fact that comorbid anxiety may lead to elevated anxiety symptoms that vary independently from depression and have probabilities of endorsement that do not exclusively depend on depression severity.

Finally, DIF was found for the items 'reactivity of mood', 'quality of mood', and 'self outlook'. These symptoms were more stable for patients in secondary care (vs. primary care) and patients with a MDD or anxiety diagnosis (vs. no diagnosis). This means that compared to diagnosed patients, for undiagnosed individuals these items are more indicative for depression at higher levels of depression. These findings fit in with prior observations. Patients diagnosed MDD have been shown to less frequently express changes in mood[40] and to have more stable negative cognitions about themselves[43]. Cognitive vulnerabilities may result in elevated symptoms at low levels of depression but these symptoms may remain stable at higher levels of depression[44], whereas patients lacking predispositions are more likely to report increased symptom levels when the burden of depression increases.

AGGREGATE-LEVEL DIF

The depression scores were adjusted for the observed DIF using methods from IRT and compared to unadjusted scores. Results indicated that DIF had a minimal effect on the total depression scores.

After adjustment for DIF, a salient change in aggregate score was only seen in a minority of the participants, (0.2%-12.7%). This fraction was largest for chronic illness (12.7%), probably because here DIF was uniform and unidirectional. Chronically ill patients were more likely to report somatic symptoms at all levels of depression, whereas for the other clinical groups, DIF was non-uniform and in opposing directions. For example, primary care patients were more likely to report somatic complaints but less likely to report anxiety related symptoms. These opposite effects cancelled each other out on the aggregate level. Another reason for the negligible effect on aggregate level is that most symptom-level DIF was observed at low and high levels of depression, whereas most patients reported medium levels of depression. Based on the current findings, the effects of DIF on depressive symptom reporting seem of little clinical relevance.

Still, the impact of DIF on depression severity measures should not be completely dismissed. There are two cases where DIF may be relevant. First, research in which the main goal is to compare depression severity between clinical groups may benefit from adjusting for DIF to improve the accuracy of the assessment, since a small percentage (4.1%-14.6%) of the differences between clinical groups is attributable to DIF. Second, observed DIF may be an indication of the presence of secondary dimensions affecting the probabilities of response. These secondary dimensions may have a more pronounced effect in samples that are more homogenous than the current sample, as one group consistently shows different probabilities of reporting symptoms. For example, patients with cardiovascular disease may have a consistently increased probability of endorsing somatic complaints and anxiety-related symptoms when filling in depression scales, leading to consistent, unidirectional depression severity overestimation. In addition, the current results suggest that DIF may have more pronounced effects in psychiatric inpatients at the severe end of the spectrum. Unfortunately, the clinical relevance of DIF in this stratum could not be assessed.

STRENGTHS AND LIMITATIONS

Amongst the strengths of this study are that we were able to investigate differences in depressive symptom reporting across distinct relevant clinical populations in a large representative sample. Furthermore, an advanced and flexible hybrid method using logistic regression and IRT was used enabling investigation of both uniform and non-uniform

DIF. Finally, clinical relevance of DIF findings was emphasized. Still, results of the present study should be interpreted with the following limitations in mind. First, results may be population-specific, especially with respect to healthcare setting, which differ across countries[45]. Second, results may be specific for the IDS-SR, as the IDS-SR contains more associated symptoms of depression than most other questionnaires. Third, conclusions are limited to the sources of DIF that were investigated in this study. Future research could apply the current analyses to more homogeneous populations, to other instruments and could include other potential sources of DIF.

IMPLICATIONS OF CURRENT STUDY

In conclusion, differences in depressive symptom reporting between clinical groups were observed on the symptom level and may be studied further to increase understanding about the phenomenology of depression. On the aggregate level these effects had minimal clinically relevant consequences, suggesting that the IDS-SR total score is robust against DIF and can be safely used by clinicians and researchers to compare depression severity across clinical groups. Based on these findings, we conclude that differences in depressive symptom reporting across clinical populations are a fact, but do not make any difference from a clinical point of view.

REFERENCES

1. Mellenbergh, G. J. Item bias and item response theory. *Int. J. Educ. Res.* **13,** 127–143 (1989).
2. Reise, S. P. & Waller, N. G. Item Response Theory and Clinical Measurement. *Annu. Rev. Clin. Psychol.* **5,** 27–48 (2009).
3. Uebelacker, L. A., Wang, P. S., Berglund, P. & Kessler, R. C. Clinical differences among patients treated for mental health problems in general medical and specialty mental health settings in the National Comorbidity Survey Replication. *Gen. Hosp. Psychiatry* **28,** 387–395 (2006).
4. Yang, F. M. & Jones, R. N. Measurement Differences in Depression: Chronic Health-Related and Sociodemographic Effects in Older Americans. *Psychosom. Med.* **70,** 993–1004 (2008).
5. Teresi, J. A., Ramirez, M., Lai, J.-S. & Silver, S. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychol. Sci. Q.* **50,** 538 (2008).
6. Reise, S. P. & Waller, N. G. How many IRT parameters does it take to model psychopathology items? *Psychol. Methods* **8,** 164–184 (2003).
7. Sudman, S., Bradburn, N. M. & Schwarz, N. *Thinking about answers: The application of cognitive processes to survey methodology*. **xiv,** (Jossey-Bass, 1996).
8. Holland, P. W. & Wainer, H. *Differential Item Functioning*. (Routledge, 1993).
9. Scott, N. W. *et al.* Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health Qual. Life Outcomes* **8,** 81 (2010).
10. Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S. & Crane, P. K. Modifying Measures Based on Differential Item Functioning (DIF) Impact Analyses. *J. Aging Health* **24,** 1044–1076 (2012).
11. Broekman, B. F. P. *et al.* Differential item functioning of the Geriatric Depression Scale in an Asian population. *J. Affect. Disord.* **108,** 285–290 (2008).
12. Osborne, R. H., Elsworth, G. R., Sprangers, M. a. G., Oort, F. J. & Hopper, J. L. The value of the Hospital Anxiety and Depression Scale (HADS) for comparing women with early onset breast cancer with population-based reference women. *Qual. Life Res.* **13,** 191–206 (2004).
13. Cole, S. R., Kawachi, I., Maller, S. J. & Berkman, L. F. Test of item-response bias in the CES-D scale: experience from the New Haven EPESE Study. *J. Clin. Epidemiol.* **53,** 285–289 (2000).
14. van Beek, Y., Hessen, D. J., Hutteman, R., Verhulp, E. E. & van Leuven, M. Age and gender differences in depression across adolescence: real or 'bias'? *J. Child Psychol. Psychiatry* **53,** 973–985 (2012).
15. Penninx, B. W. J. H. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17,** 121–140 (2008).
16. Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B. & Trivedi, M. H. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol. Med.* **26,** 477–486 (1996).
17. Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. *impute: Imputation for microarray data.* (2013).
18. Lord, F. *Applications of Item Response Theory to Practical Testing Problems*. (Routledge, 1980).
19. Meijer, R. R. & Baneke, J. J. Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychol. Methods* **9,** 354–368 (2004).
20. Sijtsma, K. & Molenaar, I. *Introduction to nonparametric item response theory*. (SAGE, 2002).
21. Ramsay, J. *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*. (2000).
22. Molenaar, I. & Sijtsma, K. *User's manuals MSP5 for Windows. IEC ProGAMMA, Groningen.* (2000).
23. van der Ark, L. Mokken scale analysis in R. *J. Stat. Softw.* **20,** (2007).
24. Reise, S. P., Moore, T. M. & Haviland, M. G. Bifactor Models and Rotations: Exploring the Extent to Which Multidimensional Data Yield Univocal Scale Scores. *J. Pers. Assess.* **92,** 544–559 (2010).
25. Wardenaar, K. J. *et al.* The structure and dimensionality of the Inventory of Depressive Symptomatology Self Report (IDS-SR) in patients with depressive disorders and healthy controls. *J. Affect. Disord.* **125,** 146–154 (2010).

26. Hu, L. & Bentler, P. M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. Multidiscip. J.* **6,** 1–55 (1999).
27. McDonald, R. *Test theory a unified treatment*. (L. Erlbaum Associates, Mahwah, N.J., 1999).
28. Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw.* **048,** (2012).
29. Choi, S. W., Gibbons, L. E. & Crane, P. K. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *J. Stat. Softw.* **39,** 1–30 (2011).
30. Cai, L., Toit, S. du & Thissen, D. *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. (Scientific Software International, 2011).
31. Crane, P. K. *et al.* A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Qual. Life Res.* **16,** 69 (2007).
32. Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychom. Monogr. Suppl.* **34,** 100 (1969).
33. Kim, S.-H., Cohen, A. S., Alagoz, C. & Kim, S. DIF Detection and Effect Size Measures for Polytomously Scored Items. *J. Educ. Meas.* **44,** 93–116 (2007).
34. Steinberg, L. & Thissen, D. Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychol. Methods* **11,** 402–415 (2006).
35. Kim, J. Effect of Multiple Testing Adjustment in Differential Item Functioning Detection. *Educ. Psychol. Meas.* **73,** 458–470 (2013).
36. Wyrwich, K. W., Tierney, W. M. & Wolinsky, F. D. Further Evidence Supporting an SEM-Based Criterion for Identifying Meaningful Intra-Individual Changes in Health-Related Quality of Life. *J. Clin. Epidemiol.* **52,** 861–873 (1999).
37. Revicki, D., Hays, R. D., Cella, D. & Sloan, J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J. Clin. Epidemiol.* **61,** 102–109 (2008).
38. Hanley, J. A. & McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148,** 839–843 (1983).
39. Peracchi, F. & Rossetti, C. Heterogeneity in health responses and anchoring vignettes. *Empir. Econ.* **42,** 513–538 (2012).
40. Suh, T. & Gallo, J. J. Symptom profiles of depression among general medical service users compared with speciality mental health service users. *Psychol. Med.* **27,** 1051–1063 (1997).
41. Katon, W., Sullivan, M. & Walker, E. Medical symptoms without identified pathology: relationship to psychiatric disorders, childhood and adult trauma, and personality traits. *Ann. Intern. Med.* **134,** 917–925 (2001).
42. Mineka, S., Watson, D. & Clark, and L. A. Comorbidity of Anxiety and Unipolar Mood Disorders. *Annu. Rev. Psychol.* **49,** 377–412 (1998).
43. Hollon, S. D. Cognitive Models of Depression From a Psychobiological Perspective. *Psychol. Inq.* **3,** 250–253 (1992).
44. Clark, D. A. & Beck, A. T. Cognitive theory and therapy of anxiety and depression: Convergence with neurobiological findings. *Trends Cogn. Sci.* **14,** 418–424 (2010).
45. Wang, P. S. *et al.* Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *The Lancet* **370,** 841–850 (2007).

CHAPTER 5

Why Hamilton was Right on Differential Weighting of Depressive Symptom Severity

Rob B. K. Wanders, Steve P. Reise,
Klaas J. Wardenaar, Peter de Jonge,
Rob R. Meijer

In preparation

ABSTRACT

Background. In the analyses of individual depressive symptoms the single items are assumed to validly measure a continuum reflecting symptom severity. This study aimed to investigate the validity of measuring individual symptoms in self-report questionnaires, and illustrates the use of score weights as an intuitive way to evaluate category ordering and functioning.

Methods. A large cohort of patients (n=2292) completed the Inventory of Depressive Symptomatology self-report (IDS-SR). Category score weights were derived from the nominal response model, which describe the information that each category provides about depression severity, without assuming an ordered response structure. Score weights were used to compute a weighted total score based on the model.

Results. Category functioning was found problematic for 13 out of 28 items. Of these, 11 items showed redundancy in the number of response options, where categories were indistinguishable with respect to depression severity. For five items, reversed categories were observed, where a higher category was not indicative for higher levels of depression severity.

Conclusions. For a significant number of items the associated categories did not validly reflect a continuum of depression severity and showed redundant and disordered categories. Results support the differential weighting of depressive symptoms. The nominal response model and derived category score weights can provide an intuitive way for researchers to ascertain proper item functioning prior to analysis of individual symptoms.

INTRODUCTION

Depressive symptoms vary in terms of their heritability[1], associated risk factors[2], relations to biomarkers[3], impact on impairment[4,5], and associations with other psychiatric[6,7] and medical diagnoses[8]. This has led researchers to question the assumed interchangeability of symptoms that are used to define a diagnosis of major depressive episode according to the DSM-5[9] and to advocate a shift from using traditional depression sum scores to the analysis of individual depressive symptoms[10].

Although standardized interviews (e.g. Composite International Diagnostic Interview [CIDI[11]] and Mini-International Neuropsychiatric Interview [MINI[12]]) are widely used, they were designed to establish diagnoses and not to assess detailed symptom severity. As such, they often skip symptoms when diagnostic core criteria are not met, and only assess the presence or absence of symptoms when more detailed information on symptom severity is wanted. In many cases, researchers therefore have to rely on depression questionnaires as their source of information on depressive symptom severity. In this study we investigate the validity of measuring symptom severity by means of single depression items and their associated categories in a self-report depression questionnaire. For this purpose, items were modelled as if they were nominal, allowing for the evaluation of each category's contribution to the total measurement of severity.

Most self-report questionnaires (and corresponding clinician-rated versions) are designed to measure depression severity by rating individual depressive symptoms on an ordered multi-point scale, often anchored with detailed descriptions to gauge the severity of each category (e.g. Hamilton Rating Scale for Depression [HRSD[13]], Inventory of Depressive Symptomatology [IDS-SR[14]]). The number of categories and their associated anchors are chosen in such a way that they measure the individual symptom along a severity continuum. Categories are then unit-weighted, with each higher category reflecting an increase in severity, adding one point towards the total score. The implicit assumptions made here are: a) that all symptoms are interchangeable and add equally towards the total score, b) that all symptoms can be meaningfully measured across a severity dimension, and c) that all response categories within and across items represent equal steps of symptom severity. Unfortunately, these fundamental assumptions underlying the measurement of symptom severity have received limited attention in research of depression instruments.

The potential problem of equal weighting of items was recognized long ago. For the HRSD, Hamilton[13] assigned some symptoms fewer categories (3 vs. 5) because he did not regard each symptom as equally suitable to be measured dimensionally[15]. However, many recent depression measures contain only equally weighted items. For the IDS-SR, Rush[14] argued that there was no rationale for differential weighting, and decided to measure every

symptom by four anchored categories. Despite the fact that two high-profile instruments take such different approaches, the appropriateness and usefulness of equal vs. differential weighting of depression items have not been thoroughly investigated. Santor et al.[16–18], investigated the category functioning of the HRSD by visually inspecting the endorsement rate on each category plotted against different levels of severity. They concluded that many items have problematic response categories that should not be considered to reflect severity on the assessed symptom.

For research to advance in the analysis of individual depressive symptoms, it is important that the validity of the measured symptom severity is beyond dispute. Recent renewed interests in the nominal response model (NRM[19]) showed that this model could serve as a useful tool to analyze item and category functioning when applied to self-report questionnaires that contain items existing of multiple ordered categories[20]. The NRM was originally developed to investigate the potential information that lies in the endorsement of false options in multiple-choice tests[19], with the idea that not every false option is equally bad. Since the model does not assume an ordered structure of categories where each additional category is indicative of increased severity, it provides an ideal way to investigate category ordering and functioning in symptom-severity assessments. For example, Murray et al.[21] analyzed the category functioning in a personality questionnaire (Personality Factor Questionnaire, version 5) and showed that the middle categories are often interpreted as either bottom or top categories and not consistently used to report medium levels of severity.

Model parameters that describe the relationship between items and underlying depression severity and the estimated severity scores in item response theory (IRT) models such as the NRM are not always easy to comprehend and interpret. This requires detailed statistical knowledge of the model, and for the non-trained researcher the contribution of categories and items to estimated depression severity often remains unclear. However, simple score weights associated with each category that summarize how each category should be weighted towards depression severity can be derived from the model and easily used to calculate corrected severity scores. As such, score weights provide an intuitive and useful tool to investigate category functioning and category ordering and to assess how these affect the validity of symptom-severity measurements.

The aim of the present study was to investigate the validity of measuring individual depressive symptoms in a self-report depression questionnaire in a large cohort study, and to use category score weights derived from the NRM to understand category functioning.

METHODS

PARTICIPANTS AND PROCEDURES

Data was derived from the Netherlands Study of Depression and Anxiety (NESDA), an ongoing longitudinal cohort study among 2981 adults (1979 women; age range: 18-65 years) to investigate the long-term course of depression and anxiety disorders. The rationale, objectives, and methods of the study have been described in detail elsewhere[22]. Participants were recruited from the community (19%), primary care (54%), and secondary care (27%). Baseline assessments took place between 2004 and 2007 and included a face-to-face assessment session with a trained research assistant, consisting of a standardized psychiatric and demographic interview, biomedical measurements, a blood-draw and a battery of self-report questionnaires. The Ethical Committees of all participating universities approved the protocol of the NESDA study. All participants signed informed consent.

Data for the present study came from the baseline assessment, and included only participants with a lifetime anxiety or depression diagnosis (n=2329; 78.1%). Of these, patients with more than 5 missing values on the IDS-SR were excluded, leading to a final sample of 2292 patients.

MEASURES

Depressive symptoms

The IDS-SR[14] is a self-report questionnaire that measures the severity of depressive symptoms with 30 items rated on a 4-point scale (0,1,2,3). The four categories are anchored with detailed descriptions that reflect the different severity levels of the assessed symptom. Each participant could either report 'appetite increase' or 'appetite decrease' and either 'weight increase' or 'weight decrease'. These items were therefore combined respectively into single 'appetite change' and 'weight change' items. All DSM-5 criterion symptoms of major depressive disorder (MDD) are assessed by the IDS-SR, as well as the most commonly associated symptoms (e.g. irritability, anxiety). Cutoffs for the IDS-SR have been suggested[23] that indicate different severity levels of depression: none (0-13), mild (14-25), moderate (26-38), severe (39-48) and very severe (49-84).

External variables

Socio-demographic factors gender and age were assessed at baseline. A standardized psychiatric interview (Composite International Diagnostic Interview; CIDI, WHO version 2.1) was conducted at baseline to assess the presence of lifetime and current (past six months) DSM-5 diagnoses of major depressive disorder (MDD), dysthymia, generalized anxiety disorder (GAD), social phobia, agoraphobia, and panic disorder. Of the Dutch

short adaptation of the Mood and Anxiety Symptoms Questionnaire (MASQ-D30[24]) the subscales 'lack of positive affect', 'negative affect' and 'somatic arousal' (each scale 10 items, range: 10-50) were used. Neuroticism was assessed using the Neuroticism-Extraversion-Openness Five-Factor-Inventory (NEO-FFI[25]). From the Four Dimensional Symptom Questionnaire (4DSQ[26]) the distress scale (16 items, range 0-31) was used.

STATISTICAL ANALYSES

Missing Data

Response data of patients with five or less missing responses on the IDS-SR were imputed on the item level (323 item scores, 0.50%). Missing values on external variables were imputed on the scale level. Imputation was performed using a K-nearest neighbor (KNN) search that imputes the missing values based on the scores of subjects reporting similar symptom profiles. The R package 'impute'[27] was used to perform KNN imputation.

Nominal response model

The NRM[19] is the most general IRT model that describes the relation between categories of an item (symptom) and the underlying depression severity without assuming an ordered structure[28]. Two parameters are estimated in this model for each response option of an item: an intercept and a slope parameter defining a model of the odds that someone endorses the response option[20]. The slope parameter then represents how strongly the odds change with the underlying trait (depression severity), and the intercept represent the odds at the zero trait level and reflects the relative popularity (i.e. higher intercepts are obtained if a category is often endorsed at lower levels of depression). These parameters are difficult to interpret however, leading researchers to explore alternatives (e.g. deriving category boundary discrimination parameters that inform on the distinction between categories[20]) that could provide better interpretable results. Based on recent developments a new parameterization of the NRM was proposed[28] that provides a more intuitive way to interpret the results of the NRM. The original model can be rewritten using a scoring-function formulation, with separate scoring and category coefficients[28]. The scoring coefficients directly reflect the information that each category gives on the underlying ability (depression severity). These can be multiplied together into single category score weights that reflect how a response on each category should be weighted towards the total score according to the NRM[29,30]. There are two interesting applications of score weights that are used in the current study. First, they provide us with the opportunity to investigate category functioning in an intuitive way. Second, all weighted responses can be summed into a single weighted total score that can be used as a proxy (sufficient statistics) for the more complex IRT trait score. These two applications are discussed next in more detail.

Category score weights

When computing raw total scores on self-report questionnaires, each response gets unit weighted depending on the category in which the response is given (e.g. for the IDS-SR responses are unit-weighted as either 0, 1, 2, or 3). Instead of unit weights, the scoring weights derived from the NRM give for each category the optimal weight to obtain an optimal score according to the model. Since the model does not assume ordered categories or equal distinctions between categories and items, each category can obtain a distinct score weight that deviates from the assumed ordered unit weights. These obtained scoring weights can be evaluated across and within each item and can inform us in an intuitive way on three different aspects regarding item and category functioning.

First, it enables us to evaluate the appropriateness of equal weighting and investigate to what extent differences exist between items and associated categories in terms of the information they provide about depression severity. Besides valuable information about measurement properties, this also informs us on how well the anchored descriptions represent the underlying symptom severity.

Second, the score weights provide information on whether an item may contain too many response categories. As discussed in the introduction, not every symptom may be meaningfully assessed along a continuum of symptom severity. Forcing these items on a four point scale can lead to a situation where the distinction between categories does not provide meaningful information about a true difference in severity. In this case, distinct categories will obtain similar score weights and the categories are indistinguishable with respect to depression severity.

Third, the categories of IDS-SR items are anchored with descriptions that are assumed to reflect an increasing order of severity. The score weights can be used to evaluate if the ordering of the score weights indeed reflects the same ordering as indicated by the category descriptions. If a higher category obtains a lower weight than a lower category, the categories are disordered, and the lower category is indicative of higher depression severity. For the IDS-SR, evaluation of this ordering will also inform us about the appropriateness of the chosen anchored descriptions.

In addition to the inspection of category score weights, it has been suggested that the assumption of ordered categories could be evaluated by comparing the NRM to the general partial credit model (GPCM[31]) in terms of fit and information criteria[20]. The GPCM is a special (nested) case of the NRM where the ordering of categories is constrained. If an ordered structure is appropriate, the GPCM should describe the data as well as the NRM, yet with fewer parameters. Therefore, the GPCM should obtain lower values on information criteria if an ordered structure describes the data well. This can both be done on the test as a whole (GPCM vs. NRM), as well as item-wise by investigating the fit of a mixed NRM

model where constraints are applied to the item under investigation and leaving all other items unconstrained.

Weighted total scores

One of the uses of IRT models is to obtain an estimate of the underlying ability (depression severity) associated with the person's response pattern that takes the differential relations of symptoms with depression into account. That is, for two persons with the same raw total score, but with different symptom patterns, the person that endorses a pattern consisting of relatively more symptoms that are indicative of increased depression severity (e.g. with sad mood and suicidal ideation) obtains a higher estimate of depression severity than a person that endorses a pattern with symptoms that are less indicative for depression. There are different estimation procedures (e.g. maximum likelihood) to find the most likely score on the underlying ability (depression severity) in IRT, but in all of these it always remains a mystery why exactly a certain estimate corresponds to a certain response pattern. However, the derived score weights from the NRM can be directly used to obtain a weighted score that serves as a proxy (sufficient statistic) for the more complex IRT trait score, where a single weighted score exists for each IRT score. These weighted scores can be obtained by multiplying each individual's response pattern with the corresponding category score weights and summing the resulting weighted scores. As such, the weighted scores can make the procedure of obtaining an estimate of depression severity based on IRT an explicit process.

It should be noted here that the correlation between raw total scores and more optimal IRT scores are often high ($r>0.9$), because the relative ordering of individuals' scores remains largely intact. That is, those with high raw total scores will also have high weighted scores, and vice versa. In many situations the use of either score will therefore lead to the same results and conclusions. However, there are situations where the increased precision that weighted scores can offer might be important. First, for total scores around the cutoff values on the IDS-SR, the weighted scores could provide higher diagnostic accuracy. Second, in analyses for which small effects are the rule rather than the exception (e.g. gene-environment interactions) the weighted scores may provide a more accurate measurement of depression severity.

As described above, patients with the same total score, but with different symptom patterns can obtain a differently weighted score. If such distinctions made between weighted scores at similar raw score levels truly reflect meaningful differences in depression severity, than the higher weighted scores should be associated with poorer clinical and psychological outcomes. To investigate this, subgroups were created for each raw total score level (i.e. consisting out of individuals that obtained the same total score). Separate regression

analyses were performed in each subgroup investigating the associations between the individuals' weighted scores and several external variables (lack of positive affect, negative affect, somatic arousal, distress, and neuroticism). This allowed us to evaluate whether variations of weighted scores conditional on total score were indeed associated with variations on adverse psychopathological outcomes. To gain a clear overall insight into the direction of these associations, the regression results were plotted in a forest plot. In addition, the individual regression analyses were combined as a meta-analysis pooled across severity levels to obtain an estimate of the overall effect[32].

Outline of analyses

All IRT analyses were performed with the R package 'mirt'[33]. First, item and model fit were investigated for the NRM and GPCM. A difference of ten points in BIC was taken as a practical cut-off to reflect a significant difference. In addition to the information criteria, the bootstrapped likelihood ratio test (BLRT) was computed to investigate whether a model significantly increased in model fit. Second, category score weights obtained from the NRM are used to investigate (i) the differential weighting of each item and associated categories, (ii) potential redundant categories (score weight difference between categories smaller than 0.2), and (iii) the ordering of categories. Third, weighted total scores were computed and their added value was evaluated. In addition, ROC analyses were performed to compare the diagnostic performance between weighted scores and raw total scores, using a CIDI diagnosis of current MDD as outcome. These analyses were conducted in the whole sample, and separately for patients with scores around the diagnostic cutoff (IDS-SR score between 14 and 39).

RESULTS

The 2292 participants had a mean age of 42.2 (SD: 12.6) and 67.9% was female. Of the sample, 51.4% had a diagnosis of a current DSM-5 mood disorder (MDD or dysthymia), and 55.9% had a diagnosis of a current DSM-5 anxiety disorder (GAD, social phobia, agoraphobia, or panic disorder). Item-level descriptives are shown in Table 1.

The NRM showed better overall fit than the GPCM (BIC: 150695 vs. 152562). Inspection of item-fit showed 17 items with poor fit under the GPCM, and 3 items with poor fit under the NRM (S-X2 statistics with $p<0.05$). For each item the assumption of ordered categories was assessed by comparing the NRM to a mixed NRM model with ordering constraints (GPCM) applied to each investigated item separately. For 13 out of 28 items the nominal model was to be preferred according to the information criteria (ΔBIC bigger than 10). BLRT showed only for 8 of 28 items that the GPCM significantly improved fit. These results

are a first indication that the assumption of ordered categories may not hold for each item in the IDS-SR.

TABLE 1. Item-level descriptives.

Item		Frequencies				Descriptives		
#	Label	0	1	2	3	Mean	(SD)	Item-rest ρ[1]
1	Onset insomnia	1091	426	444	331	1.0	(1.1)	0.33
2	Mid insomnia	568	691	602	431	1.4	(1.1)	0.24
3	Morning insomnia	1521	356	245	170	0.6	(1.0)	0.26
4	Hypersomnia	1392	675	160	65	0.5	(0.8)	0.17
5	Sad mood	700	980	477	135	1.0	(0.9)	0.71
6	Irritable	619	1076	477	120	1.0	(0.8)	0.59
7	Anxious	562	1033	582	115	1.1	(0.8)	0.65
8	Reactivity of mood	1438	543	250	61	0.5	(0.8)	0.59
9	Mood variation	1395	626	113	158	0.6	(0.9)	0.20
10	Quality of mood	961	565	422	344	1.1	(1.1)	0.57
11/12	Appetite change	1133	762	183	214	0.8	(1.0)	0.47
13/14	Weight change	1106	650	340	196	0.8	(1.0)	0.27
15	Concentration	678	990	470	154	1.0	(0.9)	0.65
16	Self outlook	973	630	96	593	1.1	(1.2)	0.51
17	Future outlook	510	1395	246	141	1.0	(0.8)	0.61
18	Suicidal ideation	1502	457	298	35	0.5	(0.8)	0.45
19	Involvement	1159	816	158	159	0.7	(0.9)	0.65
20	Energy	743	760	654	135	1.1	(0.9)	0.68
21	Enjoyment	1131	856	254	51	0.7	(0.8)	0.69
22	Sexual interest	1044	654	402	192	0.9	(1.0)	0.47
23	Psychomotor slowing	1499	321	409	63	0.6	(0.9)	0.57
24	Psychomotor agitation	1043	559	595	95	0.9	(0.9)	0.40
25	Somatic complaint	486	1139	535	132	1.1	(0.8)	0.46
26	Sympathetic arousal	645	1189	429	29	0.9	(0.7)	0.47
27	Panic	948	726	507	111	0.9	(0.9)	0.40
28	Gastrointestinal	1099	759	335	99	0.8	(0.9)	0.33
29	Sensitivity	616	1056	243	377	1.2	(1.0)	0.50
30	Leaden paralysis	476	938	493	385	1.3	(1.0)	0.65

[1] Correlation between the item and the total rest score (scored without item).

NOMINAL SCORING WEIGHTS

Score weights derived from the NRM for each item and associated categories are shown in Table 2 and Figure 1. The score weights differed across items and categories, and deviated strongly from the raw integer weights (1, 2, 3) in each item. The maximum weight of each symptom differed strongly across items, ranging from 0.7 for 'mood variation' and 6.9 for 'enjoyment'. This indicated that items differentially contribute to underlying depression severity with some items contributing only little information. Interestingly, for some items the weights of the first categories were already larger than the weights of the maximum category for many other symptoms. For example, nine items had maximum weights below 2.0, whereas six items already obtained a weight on the first category above 2.0 (e.g. 3.2 for 'quality of mood'). This indicated that the first categories of these items are more indicative and discriminative for depression severity than the highest categories of many other items.

REDUNDANT CATEGORIES

Possible redundant categories were observed for 11 items (Table 2, Figure 1), where two or more categories obtained comparable score weights (difference in weights smaller than 0.2). This implies that for these items the categories do not discriminate between individuals in terms of depression severity, and can therefore be considered redundant (Figure 2). For both the items 'onset insomnia' and 'hypersomnia' the second category obtained a weight near zero (0.1), and for 'mid insomnia' the third category obtained a weight near zero (0.1), suggesting that these categories provide little information on depression severity and behave as the first category (absence of symptom). For 'weight change' the two middle categories obtained similar weights (0.5 vs 0.6) and for the items 'appetite change', 'gastrointestinal', and 'leaden paralysis' the last category had a score weight that was similar to the weight of the upper middle category. This means that for these items two categories function almost the same. An illustration in Figure 2 shows the category curves associated with the item 'gastrointestinal'. Here, it can be seen that the probability of giving a response in category 3 or 4 is the same regardless of depression severity. There were three items 'morning insomnia', 'quality of mood', and 'psychomotor agitation' that obtained similar score weights for all three non-zero categories (categories 2, 3, and 4). For these items there is little discrimination across categories, and the items behave more like a dichotomous item. That is, only the presence of the symptom seems to be informative for depression severity with the different categories giving no additional information about severity variations.

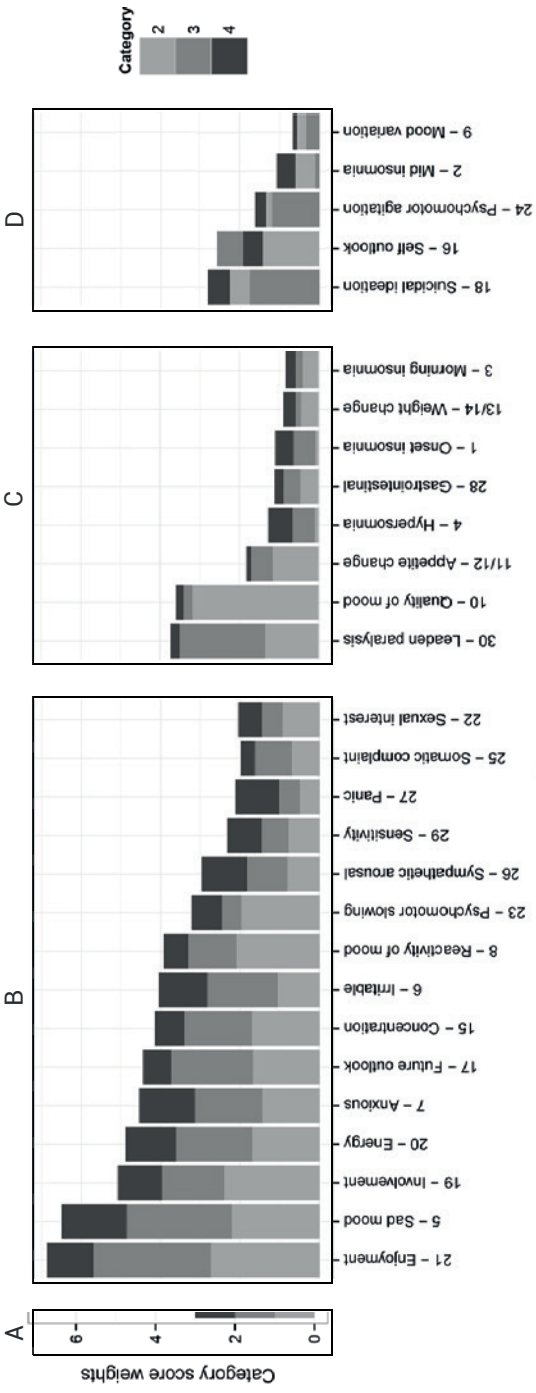


FIGURE 1. Category score weights for raw integer scoring (A), for items with good functioning categories (B), for items with redundant categories (C), and for items with disordered categories (D).

TABLE 2. Category functioning of IDS-SR items.

Item[1]		Score weights[2]				Nominal vs. ordered[3]			
#	Label	1	2	3	4	ΔAIC	ΔBIC	-2LL	BLRT p
1	Onset insomnia	0	0.1	0.6	1.1	16.0	4.5	20.0	0.000
2	Mid insomnia	0	0.6	0.1	1.1	161.3	149.9	165.3	0.000
3	Morning insomnia	0	0.4	0.6	0.8	-0.5	-12.0	3.5	0.172
4	Hypersomnia	0	0.1	0.7	1.3	19.3	7.8	23.3	0.000
5	Sad mood	0	2.2	4.9	6.5	8.0	-3.5	12.0	0.001
6	Irritable	0	1.0	2.8	4.0	24.1	12.6	28.1	0.000
7	Anxious	0	1.4	3.1	4.5	0.4	-11.1	4.4	0.118
8	Reactivity of mood	0	2.1	3.3	3.9	36.1	24.6	40.1	0.000
9	Mood variation	0	0.6	0.3	0.7	38.2	26.8	42.2	0.000
10	Quality of mood	0	3.2	3.4	3.6	400.8	389.4	404.8	0.000
11/12	Appetite change	0	1.2	1.7	1.8	72.7	61.2	76.7	0.000
13/14	Weight change	0	0.5	0.6	0.9	5.3	-6.1	9.3	0.011
15	Concentration	0	1.7	3.4	4.1	21.8	10.3	25.8	0.000
16	Self outlook	0	1.7	2.8	2.2	234.6	223.1	238.6	0.000
17	Future outlook	0	1.7	3.7	4.4	21.2	9.7	25.2	0.000
18	Suicidal ideation	0	1.9	1.4	2.6	203.4	191.9	207.4	0.000
19	Involvement	0	2.4	4.0	5.1	30.2	18.7	34.2	0.000
20	Energy	0	1.7	3.6	4.9	4.1	-7.4	8.1	0.029
21	Enjoyment	0	2.7	5.7	6.9	13.8	2.3	17.8	0.000
22	Sexual interest	0	0.9	1.4	2.0	10.0	-1.5	14.0	0.001
23	Psychomotor slowing	0	2.0	2.5	3.2	56.3	44.8	60.3	0.000
24	Psychomotor agitation	0	1.3	1.2	1.6	141.0	129.6	145.0	0.000
25	Somatic complaint	0	0.7	1.6	2.0	8.7	-2.8	12.7	0.005
26	Sympathetic arousal	0	0.8	1.8	3.0	1.5	-10.0	5.5	0.064
27	Panic	0	0.5	1.0	2.1	11.5	0.0	15.5	0.000
28	Gastrointestinal	0	0.5	0.9	1.1	-1.5	-13.0	2.5	0.313
29	Sensitivity	0	0.8	1.5	2.3	-3.1	-14.6	0.9	0.650
30	Leaden paralysis	0	1.4	3.5	3.7	104.4	92.9	108.4	0.000

AIC=Akaike Information Criteria; BIC=Bayesian Information Criteria; LL=Log Likelihood; BLRT=Bootstrapped Likelihood Ratio Test.

[1] Items with disordered categories in bold.

[2] Possible redundant categories underlined (difference in score weight smaller than 0.2)

[3] Differences in information criteria are reported, with positive difference favoring nominal categories (Nominal Response Model) instead of ordered categories (General Partial Credit Model) for that item.

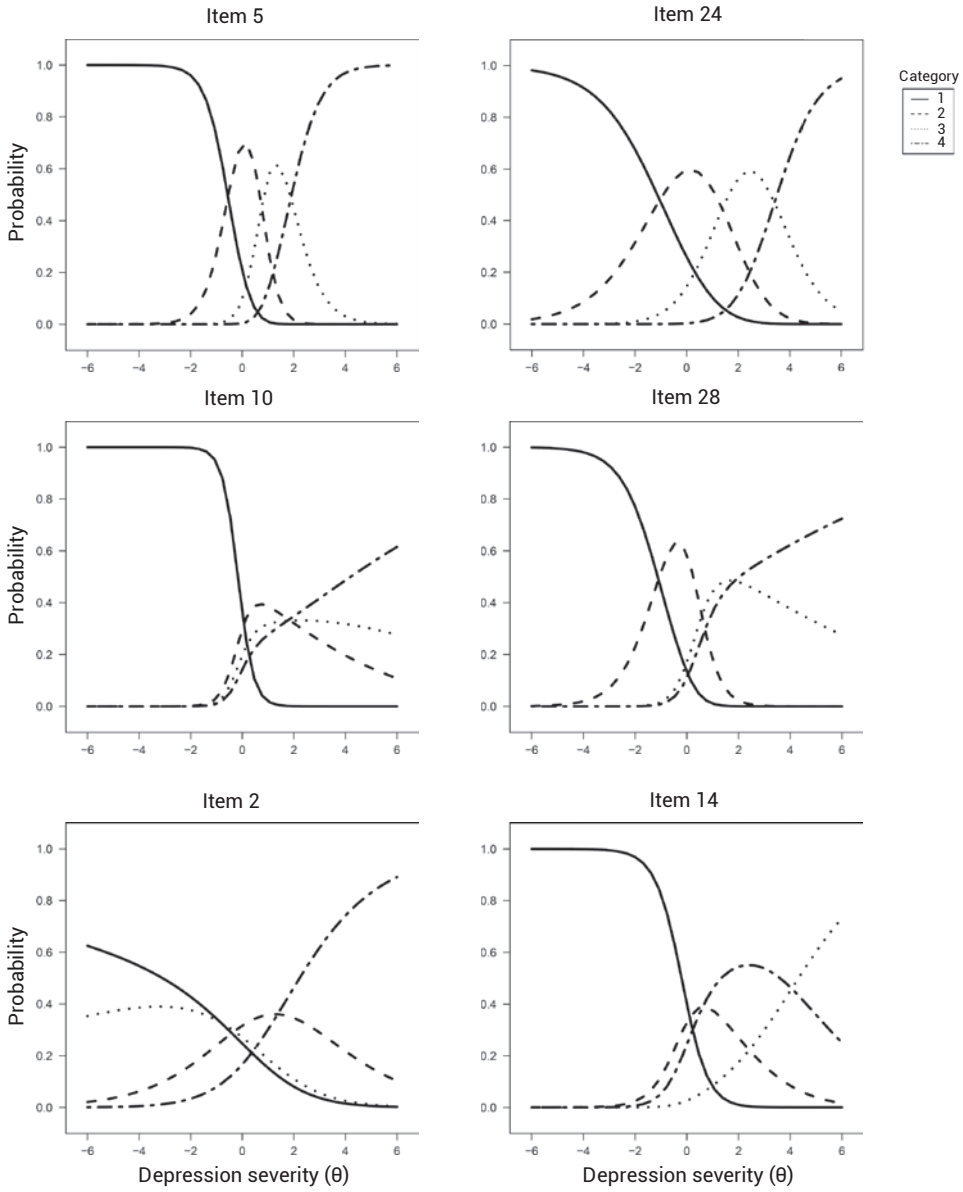


FIGURE 2. Plots of category functioning for good items (top), items with redundant categories (middle) and items with disordered categories (bottom).

DISORDERED CATEGORIES

Five items showed disordered categories (Table 2, Figure 1). Four of these items also showed very large differences in BIC when comparing fit of the NRM to the GPCM, suggesting that an ordinal response scale does not apply to these items. For the items 'mid insomnia', 'mood variation', 'suicidal ideation', and 'psychomotor agitation' the middle categories showed a reversed order. For these items the second category was indicative of higher depression severity than the third category. For the item 'self-outlook' the highest category did not get the highest weight. Besides the fact that the categories were not ordered correctly, the items also showed similar score weights across multiple categories. Figure 2 illustrates this: the category curves for the item 'mid insomnia' show that the curve belonging to the third category is positioned at a lower levels of depression severity and functions much like the first category (zero: absence of the symptom). Indistinguishable categories were also observed for 'mood variation' and 'psychomotor agitation'. This furthermore suggests that for these items the categories do not correctly reflect a continuum of depression severity. These results will be addressed in more detail in the discussion section of this paper.

WEIGHTED SCORES

The weighted scores provided a sufficient statistic for the estimated IRT trait scores obtained from the NRM, where for each IRT trait score a single weighted score exists (Figure 3). For each total score different weighted scores were obtained due to differential weighting of symptom patterns, causing variation in weighted scores conditional on total score levels (Figure 3). The correlation between weighted scores and raw total scores was high (r=0.97). A small improvement in internal consistency was observed, with a Cronbach's alpha of 0.90 (95% CI: 0.90-0.91) for the raw scores, and 0.92 (95% CI: 0.92-0.92) for weighted scores. ROC analyses also showed a small improvement in diagnostic accuracy for the weighted scores (AUC=0.83 vs AUC=0.80; p<0.001; DeLong's test) in the overall sample. Differences were greater in a subsample of patients with IDS-SR scores in the region of mild to moderate depression (IDS-SR score between 14 and 39; n=1337; 672 patients with current CIDI diagnosis of MDD), with a correlation of 0.91 between weighted and raw total scores, and a slightly bigger improvement in diagnostic accuracy (AUC=0.73 for weighted scores, and AUC=0.68 for raw scores; p<0.001; DeLong's test).

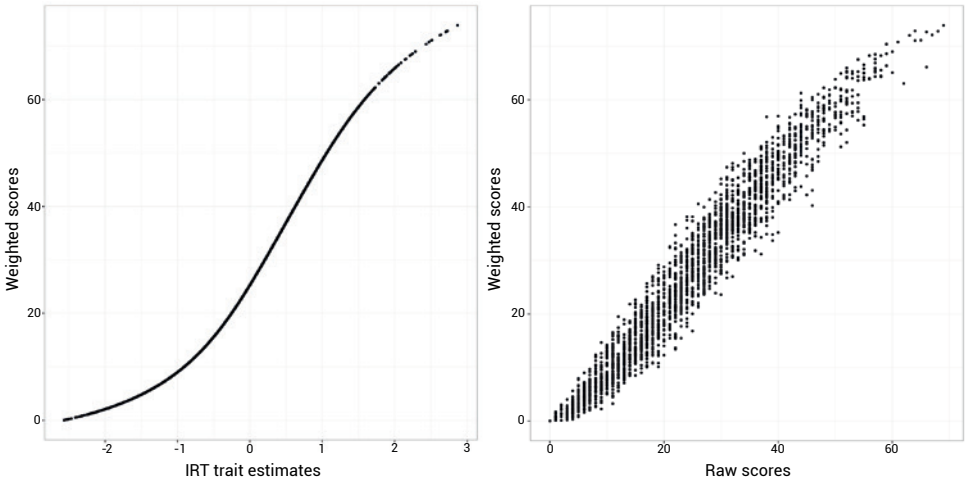


FIGURE 3. Weighted scores plotted against IRT trait estimates from the NRM (left) and against raw total scores (right).

To investigate whether weighted scores provide any additional information on depression severity compared to raw total scores, multiple regression analyses were performed in subgroups of individuals with the same raw total score but with different weighted scores, depending on their symptom pattern (Figure 4). Associations of weighted scores conditional on total score levels with external variables were predominantly in the expected direction and predicted poorer scores for 'lack of positive affect', 'negative affect', 'distress', and 'neuroticism'. For 'somatic arousal' no consistent association was observed, suggesting that the weighted scores mainly provide additional information about individuals' levels of psychological distress and affect.

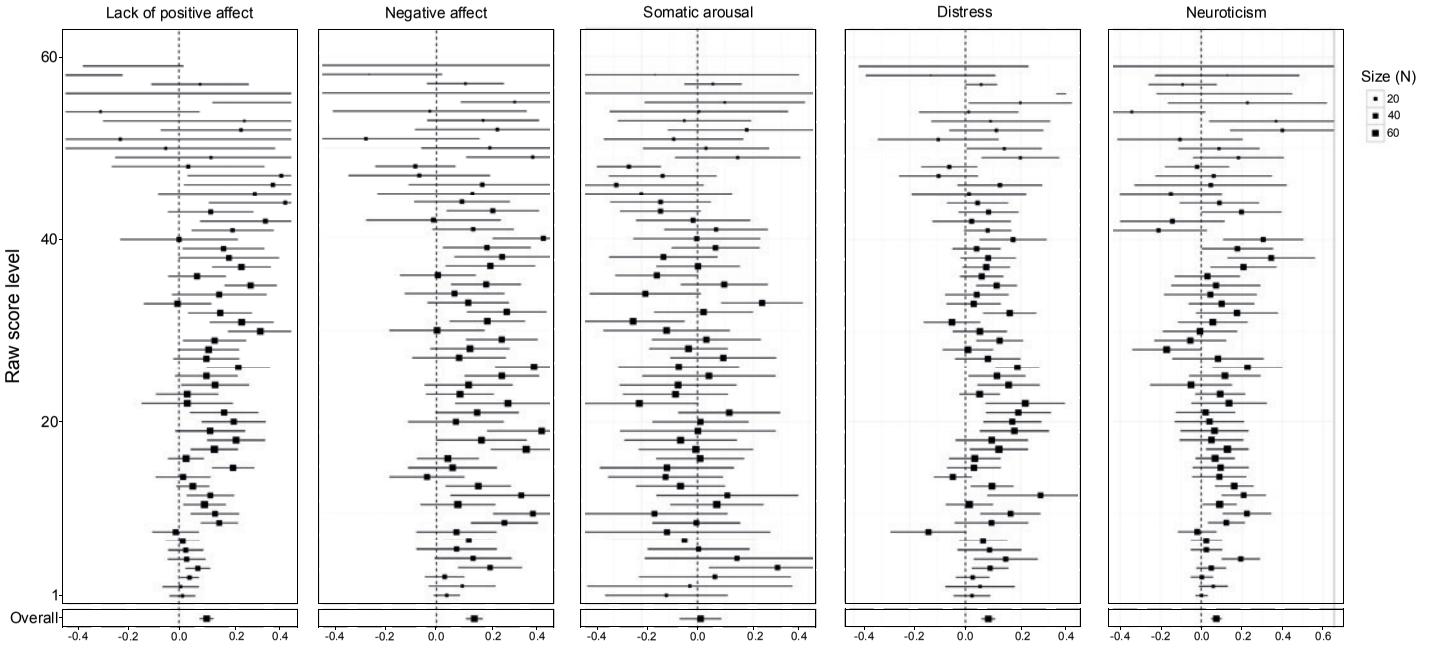


FIGURE 4. Associations between weighted scores and external variables within each raw score level.

DISCUSSION

The recent shift in focus from traditional analysis of total scores towards the analysis of individual depressive symptoms raises new questions about the validity of measuring symptom severity with single items in existing questionnaires. The current study of the IDS-SR showed that the NRM can provide very useful insights into item and category functioning and revealed several potential problems with the way the items measure a continuum of symptom severity. The observed problems and their implications are discussed below in a point-by-point fashion.

The analyses revealed eleven items for which redundant categories were observed. For these, the categories were indistinguishable with respect to depression severity, suggesting that the item could just as well be described with fewer categories. In two extreme cases all three categories measuring the presence of the symptom (categories 1,2, and 3) obtained equal score weight: 'morning insomnia' (0 vs. 0.4,0.6,0.8) and 'quality of mood' (0 vs. 3.2,3.4,3.6). For these symptoms it is even questionable whether the categories of the items provide any information on top of the dichotomy between the presence vs. absence of the symptom. The results support the suspicion of Hamilton[15] that it might not be possible to measure each symptom meaningful with multiple categories. This is in line with recent ideas that suggest that symptoms are not simply interchangeable, as is assumed in the approach underlying the current diagnostic criteria[9].

The four sleep items on the IDS-SR are known to function differently from the rest of the questionnaire[34,35], which is supported by the results of the current study. The items showed problematic category functioning with many redundant categories and a disordering of categories for 'mid insomnia'. The score weights overall were quite low (range 0-1.3) along with low item-rest correlations (0.17-0.33), suggesting that the items provide relatively little information on depression severity. There were many categories obtaining similar score weights that indicate that the categories do not make any distinction regarding differences in severity. For both 'onset insomnia' and 'hypersomnia' the second category obtained 0.1 weight, suggesting that these are equally indicative for depression as the absence of both symptoms. An explanation for this poor functioning is that sleep disturbances are common in the general population, and this may lead to many responses in the categories that are reported for other reasons than depression. Overall, these results raise the question if the four sleep items validly measure symptom severity on a four-point scale, as anchored now in the IDS-SR. In the 16-item version of the IDS-SR (QIDS-SR[36]) these sleep items are combined into a single item by taking the maximum response on all four items. Although this does not solve the disordering of the middle categories on 'mid

insomnia', such a compound item might suffice to capture the relevant information while getting rid of variance introduced by redundant response options. Sleep disturbances are a very important part of depression, but the current results suggest that the IDS-SR sleep items and associated categories function poorly. Therefore, researchers should be careful when interpreting these items' scores as they may not reflect continuous symptom severity.

The disordering of categories is a fundamental problem[37], and exposes issues in the data where endorsing a higher response category is not necessarily reflective of higher severity. The current study identified five IDS-SR items with categories that did not follow the assumed ordered structure. Investigating these results together with the anchored category descriptions of each item suggests three different explanations.

First, looking at the anchored descriptions of 'mid insomnia', and 'psychomotor agitation' it is possible that the ordering of severity is indeed reversed. For the item 'mid insomnia' the third category 'I wake up at least once a night, but I go back to sleep easily' functioned the same as the lowest category, and thus less severe than the second category 'I have a restless, light sleep with a few brief awakenings each night'. These descriptions may be somewhat unfortunate where one could argue that a restless night with multiple awakenings might indeed be more severe than a night where you awaken but go back to sleep easily. Similarly, for 'psychomotor agitation' the third category 'At times, I am unable to stay seated and need to pace around' was found slightly less severe than the second category 'I have impulses to move about and am quite restless'. These categories were also found to be indistinguishable in terms of their score weights. Indeed the category descriptions can be judged as rather overlapping in item content, asking basically the same thing in different words. The addition of 'at times' for the third category might cause some people to interpret it as less severe.

Second, the item measuring 'suicidal ideation' showed a disordering of the middle categories that might be explained by secondary factors that cause people to endorse the third category while not having alarming thoughts of death or suicidal ideas. The score weights suggest that the third category asking 'I think of suicide or death several times a week for several minutes' is less severe than 'I feel that life is empty or wonder if it's worth living'. This effect may be explained by the fact that people can have thoughts about death that are harmless, for example due to the confrontation of death in their family, and respond to the third category while not being depressed or suicidal. The second category that asks about an empty life may be more specific to depression. In contrast, the item 'self outlook' has the last two categories reversed, where the third category 'I largely believe that I cause problems for others' is deemed more severe than the fourth category 'I think almost constantly about major and minor defects in myself'. Self-criticism is less specific

for depression than self-blame[38], causing people with lower levels of depression also to endorse the highest category of 'self outlook'.

Third, the item 'diurnal variation' is not a straightforward item. First of all the item is the only item in the questionnaire for which people have to answer additional items when a certain response is given. Second, the item tries to measure multiple things, namely whether mood varies during the day (option 1 vs 2-3-4), whether mood varies due to environmental events (option 2), or whether mood varies more with the time of the day then due to environmental events (option 3). The ambiguities within this item might, on the one hand, cause people to be unsure what to respond, and on the other hand, might cause the distinction between categories to not be a unequivocal reflection of symptom severity. The disordering of the middle categories and the equal functioning of the second and last categories might suggest that only the dichotomous distinction of whether mood varies during the day (zero vs. the rest) is informative for depression severity.

The score weights were also used to obtain weighted total scores that served as a proxy for the more complex IRT trait score. The weighted scores are not meant as a way to correct for disordered categories. Violations of ordering and undiscriminating categories are signs that there are serious problems in the data which are also reflected in the weighted scores. Instead, the weighted scores were computed to illustrate how the category score weights could be used to obtain a weighted score and potential advantages were evaluated. As far as we are aware, this is the first time that a sufficient statistic based on score weights is used for the NRM. Instead of technical estimation procedures to obtain an IRT trait score, the score weights offer an intuitive way to obtain a weighted score that can be used and interpreted in a way that is analogous to raw total scores. In addition, the weighted scores make the IRT process of estimating the underlying ability (depression severity) explicit. The weighted scores correlated very highly (0.97) with the raw total scores, and only a minor improvement of internal consistency and diagnostic accuracy was observed when using weighted instead of raw scores. This suggests that using the weighted scores makes minimal practical difference and confirms that there is little change in the relative ordering of patients when switching from raw to IRT-based severity scores. This is reassuring for those using IDS-SR severity scores in clinical depression assessment and monitoring, and in line with previous studies showing the robustness of total scores[39]. However, additional analyses showed that for those with scores around the diagnostic cutoff, the difference between weighted scores and raw scores in diagnostic accuracy was more pronounced. In addition, within groups of patients that reported the same raw IDS-SR total score, higher weighted scores were associated with adverse psychopathological outcomes, suggesting that the weighted scores provide additional information about severity variations among those with seemingly similar severity judged by their raw scores. In depression research

effect sizes are often small (e.g. genetic associations[40]) and increasing measurement precision may offer a way to capture valid and more relevant phenotypic variation[41]. Indeed, previous studies have shown that IRT scores decrease the chance of finding spurious interactions[42] and improves the accuracy of analyses in gene-environment studies[43].

STRENGTHS AND LIMITATIONS

This is the first study that takes advantage of new developments in NRM parameterization, allowing for the derivation of score weights that summarize the information on category functioning provided by the NRM. In addition to the used methodology, strengths of the study included the size of the sample and the availability of relevant external measures.

There were however also limitations. First, results of the current study may be limited to the IDS-SR and to the Dutch translation used in the current study. However, results are in line with previous studies looking at category functioning in the HRSD[18] and the Dutch IDS-SR has psychometric properties similar to the English version. Second, the NRM model assumes that there exists an underlying trait, and as such, conclusions on individual depressive symptom severity are limited in as far that they are assumed to be part of a set of items that together measure depression severity. Recently, authors have opposed the idea of an underlying trait and suggested that psychopathology emerges from the dynamic interplay between symptoms over time[9]. However, irrespective of one's view of what exactly constitutes psychopathology (e.g. trait vs. network), the severity on an individual symptom is correlated with a person's overall severity level, or else the use of an ordinal response scale would make little sense. Therefore, symptoms should still have properly ordered and discriminating categories. Disordered categories are always a sign that something is wrong, and undiscriminating redundant categories should always be investigated closer.

CONCLUSION

With the focus of analyses more on individual symptoms, these results stress the importance of validity on a smaller level than overall test properties. We do not feel that the questionnaire is useless to study individual symptoms, but researchers should be careful when analyzing individual symptoms from a questionnaire that was not validated to do so. The category score weights derived from the nominal response model provide an intuitive way for researchers to ascertain proper item functioning prior to analysis of individual symptoms.

REFERENCES

1. Jang, K. L., Livesley, W. J., Taylor, S., Stein, M. B. & Moon, E. C. Heritability of individual depressive symptoms. *J. Affect. Disord.* **80,** 125–133 (2004).
2. Fried, E. I., Nesse, R. M., Guille, C. & Sen, S. The differential influence of life stress on individual symptoms of depression. *Acta Psychiatr. Scand.* **131,** 465–471 (2015).
3. Duivis, H. E., Vogelzangs, N., Kupper, N., de Jonge, P. & Penninx, B. W. J. H. Differential association of somatic and cognitive symptoms of depression and anxiety with inflammation: Findings from the Netherlands Study of Depression and Anxiety (NESDA). *Psychoneuroendocrinology* **38,** 1573–1585 (2013).
4. Tweed, D. L. Depression-related impairment: estimating concurrent and lingering effects. *Psychol. Med.* **23,** 373–386 (1993).
5. Fried, E. I. & Nesse, R. M. The Impact of Individual Depressive Symptoms on Impairment of Psychosocial Functioning. *PLOS ONE* **9,** e90311 (2014).
6. Lux, V. & Kendler, K. Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria., Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychol. Med. Psychol. Med.* **40, 40,** 1679, 1679–1690 (2010).
7. Boschloo, L. *et al*. The Network Structure of Symptoms of the Diagnostic and Statistical Manual of Mental Disorders. *PLOS ONE* **10,** e0137621 (2015).
8. Hoen, P. *et al.* Differential associations between specific depressive symptoms and cardiovascular prognosis in patients with stable coronary heart disease., Differential associations between specific depressive symptoms and cardiovascular prognosis in patients with stable coronary heart disease. *J. Am. Coll. Cardiol. J. Am. Coll. Cardiol.* **56, 56,** 838, 838–844 (2010).
9. Borsboom, D. & Cramer, A. O. J. Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annu. Rev. Clin. Psychol.* **9,** 91–121 (2013).
10. Fried, E. I. & Nesse, R. M. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med.* **13,** 72 (2015).
11. Robins, L. N. *et al.* The Composite International Diagnostic Interview: An Epidemiologic Instrument Suitable for Use in Conjunction With Different Diagnostic Systems and in Different Cultures. *Arch. Gen. Psychiatry* **45,** 1069–1077 (1988).
12. Sheehan, D. V. *et al.* The Mini-International Neuropsychiatric Interview (M.I.N.I): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry* **59,** 22–33 (1998).
13. Hamilton, M. A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* **23,** 56 (1960).
14. John Rush, A. *et al.* The inventory for depressive symptomatology (IDS): Preliminary findings. *Psychiatry Res.* **18,** 65–87 (1986).
15. Bagby, R. M., Ryder, A. G., Schuller, D. R. & Marshall, M. B. The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight? *Am. J. Psychiatry* **161,** 2163–2177 (2004).
16. Santor, D. A., O, J. & Zuroff, D. C. Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychol. Assess.* **6,** 255–270 (1994).
17. Santor, D. A., Zuroff, D. C., O, J., Cervantes, P. & Palacios, J. Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychol. Assess.* **7,** 131–139 (1995).
18. Santor, D. A. & Coyne, J. C. Examining symptom expression as a function of symptom severity: Item performance on the Hamilton Rating Scale for Depression. *Psychol. Assess.* **13,** 127–139 (2001).
19. Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* **37,** 29–51 (1972).
20. Preston, K. & Reise, S. in *Handbook of Item Response Theory Modelling: Applications to typical performance assessment* 386–405 (Routledge/Taylor & Francis Group, 2014).
21. Murray, A. L., Booth, T. & Molenaar, D. When Middle Really Means 'Top' or 'Bottom': An Analysis of the 16PF5 Using Bock's Nominal Response Model. *J. Pers. Assess.* **98,** 319–331 (2016).

22. Penninx, B. W. J. H. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17,** 121–140 (2008).
23. Rush, A. J. *et al.* The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol. Psychiatry* **54,** 573–583 (2003).
24. Wardenaar, K. J. *et al.* Development and validation of a 30-item short adaptation of the Mood and Anxiety Symptoms Questionnaire (MASQ). *Psychiatry Res.* **179,** 101–106 (2010).
25. Costa Jr, P. & McCrae, R. *Neo personality inventory–revised (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual*. (Psychological Assessment Resources, 1992).
26. Terluin, B. *et al.* The Four-Dimensional Symptom Questionnaire (4DSQ): a validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization. *BMC Psychiatry* **6,** 34 (2006).
27. Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. *impute: Imputation for microarray data*. (2013).
28. Thissen, D., Cai, L. & Bock, R. in *Handbook of polytomous item response theory models* 43–75 (2010).
29. Andrich, D. A rating formulation for ordered response categories. *Psychometrika* **43,** 561–573 (1978).
30. Ostini, R. & Nering, M. *Ostini, R., & Nering, M. L. (2006). Polytomous item response theory models (No. 144). Sage.* (Sage, 2006).
31. Muraki, E. A Generalized Partial Credit Model: Application of an Em Algorithm. *ETS Res. Rep. Ser.* **1992,** i-30 (1992).
32. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36,** 1–48 (2010).
33. Chalmers, R. mirt: A multidimensional item response theory package for the R environment. *J. Stat. Softw.* **48,** 1–29 (2012).
34. Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B. & Trivedi, M. H. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol. Med.* **26,** 477–486 (1996).
35. Wardenaar, K. J. *et al.* The structure and dimensionality of the Inventory of Depressive Symptomatology Self Report (IDS-SR) in patients with depressive disorders and healthy controls. *J. Affect. Disord.* **125,** 146–154 (2010).
36. Trivedi, M. H. *et al*. The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychol. Med.* **34,** 73–82 (2004).
37. Andrich, D. in *Handbook of polytomous item response theory models* 123–152 (2010).
38. Kannan, D. & Levitt, H. M. A review of client self-criticism in psychotherapy. *J. Psychother. Integr.* **23,** 166–178 (2013).
39. Wanders, R. B. K. *et al.* Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make? *J. Psychosom. Res.* **78,** 130–136 (2015).
40. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48,** 624–633 (2016).
41. Reise, S. & Rodriguez, A. Item response theory and the measurement of psychiatric constructs: some empirical and conceptual issues and challenges. *Psychol. Med.* **46,** 2025–2039 (2016).
42. Kang, S.-M. & Waller, N. G. Moderated Multiple Regression, Spurious Interaction Effects, and IRT. *Appl. Psychol. Meas.* **29,** 87–105 (2005).
43. Murray, A. L., Molenaar, D., Johnson, W. & Krueger, R. F. Dependence of Gene-by-Environment Interactions (GxE) on Scaling: Comparing the Use of Sum Scores, Transformed Sum Scores and IRT Scores for the Phenotype in Tests of GxE. *Behav. Genet.* **46,** 552–572 (2016).

CHAPTER 6

Problems with Latent Class Analysis to Detect Data-driven Subtypes of Depression

Hanna M. van Loo, Rob B. K. Wanders,
Klaas J. Wardenaar, Eiko I. Fried

Molecular Psychiatry 2016

Depressed patients differ considerably with respect to symptom profiles, course of illness and treatment response. These differences likely contribute to the on average low efficacy of treatment, and drive the search for more homogenous subtypes of depression in order to facilitate treatment decisions in clinical practice[1]. Latent class analysis (LCA) presents a common statistical method in current depression research that aims to identify depressed patients with similar symptom profiles[2–4]. LCA recovers hidden groups in multivariate data of heterogenous populations such that subjects within classes are similar to each other but different form subjects in other classes. It does so by dividing subjects into groups for which the observed variables are unrelated within each class, so-called 'conditional independence'[5]. Given the heterogeneity and multifactorial nature of depression[6], LCA and other multivariate subtyping strategies may yield subtypes with a more homogenous etiology, course of illness or treatment response, than subtyping depressed patients purely on one characteristic, such as with or without anxiety or psychotic features.

In a recent report in Molecular Psychiatry, Milaneschi et al.[7] used LCA and identified three classes described as 'severe typical' (T), 'severe atypical' (A) and 'moderate'. The two depressed classes T and A differed predominantly with regard to appetite and weight symptoms: most T subjects reported appetite and weight decrease, but almost none reported appetite or weight gain; for A subjects, it was the other way around. Importantly, T and A subtypes did not differ substantially with respect to other depressive symptoms, illustrated by the fact that increased appetite/weight perfectly predicted membership in A (area under the receiver operating characteristic curve (area under curve)=0.99, sensitivity 98.4%, specificity 99.5%), and decreased appetite/weight predicted membership in T very well (area under curve=0.81, sensitivity 87.8%, specificity 72.8%). Both T and A classes are consistent with results from prior LCA-based depression studies[2,3,8].

We commend the authors for their insightful study with important findings concerning the genetic background of depression, in particular that severe depression—especially when it involves appetite and weight loss (T class)—shares genetic risk factors with schizophrenia. Milaneschi et al. also demonstrated that results from multivariate classification procedures such as LCA can be used to derive more parsimonious subtypes that could serve as an alternative in case complete symptom data are unavailable (for example, in case of missing data in combined genome-wide association study data sets), which would complicate the application of classical LCA[9], as well as other multivariate subtyping techniques that we advocate below. However, we see several difficulties with the LCA-results and their interpretation that are common in the literature and not limited to the report by Milaneschi et al.[7]

First, the symptom profiles of T and A were remarkably similar and mainly differed regarding appetite/weight loss or gain. This implies that substantive variability is likely to remain among patients within these two classes with regard to other symptoms, etiology, course and prognosis, raising concerns about the value of the identified classes as means to effectively decrease the heterogeneity of depression. Validation studies are needed to test whether the T and A subtypes, despite their relatively similar symptom profiles, are differentially associated with clinically relevant external variables such as course of illness, family history or treatment outcome.

The second point pertains to the validity of these classes. Like prior reports[2,3,8], LCA classes were primarily based on weight/appetite differences that possibly reflect methodological artifacts based on violations of conditional independence. In LCA, associations between symptoms are assumed to be explained exclusively by their relation with the underlying depression subtype: symptoms within classes are statistically independent, conditional on class membership[10–12]. However, appetite/weight gain excludes appetite/weight loss in most patients (and vice versa), making these symptom-variants inherently dependent. High levels of dependence might exist as well for other opposite depressive symptoms, such as insomnia versus hypersomnia and psychomotor agitation versus psychomotor retardation. In such cases, local independence can always be achieved by increasing the number of LCA classes to account for this dependence, for instance with appetite/weight gainers allocated to a different class than appetite/weight losers[10]. The strong dependence between weight and appetite symptoms can therefore dominate the model and lead to biased parameters and posterior classifications as well as artificial classes.

Several solutions exist to account for this problem of local dependence, such as local dependence models or using Bayesian priors in so-called 'flexible LCA'[12,13]. A recent study applied both LCA and flexible LCA to depression data; regular LCA identified weight/appetite-based classes, whereas these classes disappeared in flexible LCA, which found classes differing primarily on anxiety[4]. The results emphasize the possible methodological artificiality of appetite/weight based LCA classes. Controlling for violations of conditional independence and analyzing common symptoms beyond the Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria for depression (like anxiety) may provide important venues for future research.

Third, the authors labeled the class with increased appetite/weight as 'atypical', which is custom in studies with similar results[2,8]. However, the symptom profile of this LCA-class differs considerably from the atypical specifier in the DSM, which includes additional criteria such as hypersomnia, mood reactivity, leaden paralysis and interpersonal rejection

sensitivity. Using the term 'atypical' for a class mainly characterized by increased appetite and weight might lead to further confusion in the already conflicting and contentious literature on subtypes of major depression, in which labels such as 'atypical' are used in different contexts for different combinations of criteria[1]. To prevent confusion, we suggest to use different labels for latent classes if there is no substantial overlap with specifiers used in the DSM.

Lastly, LCA assumes that classes differ only qualitatively, contrasting evidence that depression may be dimensional for some people[11]. Hybrid factor mixture models combine aspects of both LCA and factor models, allowing for the identification of classes that differ both in terms of qualitative and quantitative aspects[14]. Since the classical LCA studies, like the study by Milaneschi et al.[7], already showed promising results, addressing the abovementioned challenges will further benefit the search for empirically based depression subtypes.

REFERENCES

1. Baumeister, H. & Parker, G. Meta-review of depressive subtyping models. *J. Affect. Disord.* **139,** 126–140 (2012).
2. Lamers, F. *et al.* Identifying Depressive Subtypes in a Large Cohort Study: Results From the Netherlands Study of Depression and Anxiety (NESDA). *J. Clin. Psychiatry* **71,** 1582–1589 (2010).
3. Ulbricht, C. M., Dumenci, L., Rothschild, A. J. & Lapane, K. L. Changes in depression subtypes for women during treatment with citalopram: a latent transition analysis. *Arch. Womens Ment. Health* **19,** 769–778 (2016).
4. Have, M. ten *et al.* The identification of symptom-based subtypes of depression: A nationally representative cohort study. *J. Affect. Disord.* **190,** 395–406 (2016).
5. Oberski, D. in *Modern Statistical Methods for HCI* (eds. Robertson, J. & Kaptein, M.) 275–287 (Springer International Publishing, 2016).
6. Kendler, K. & Prescott, C. *Genes, environment and psychopathology : understanding the causes of psychiatric and substance use disorders.* (Guilford Press, 2006).
7. Milaneschi, Y. *et al.* Polygenic dissection of major depression clinical heterogeneity. *Mol. Psychiatry* **21,** 516–522 (2016).
8. Sullivan, P. F., Kessler, R. C. & Kendler, K. S. Latent Class Analysis of Lifetime Depressive Symptoms in the National Comorbidity Survey. *Am. J. Psychiatry* **155,** 1398–1406 (1998).
9. Suppes, P. & Zanotti, M. When are probabilistic explanations possible? *Synthese* **48,** 191–9 (1981).
10. Borsboom, D. *et al.* Kinds versus continua: a review of psychometric approaches to uncover the structure of psychiatric constructs. *Psychol. Med.* **46,** 1567–1579 (2016).
11. Hagenaars, J. A. Latent Structure Models with Direct Effects between Indicators. *Sociol. Methods Res.* **16,** 379–405 (1988).
12. Asparouhov, T. & Muthén, B. in *Proceedings of the 2011 Joint Statistical Meeting* (2011).
13. Miettunen, J., Nordström, T., Kaakinen, M. & Ahmed, A. O. Latent variable mixture modeling in psychiatric research – a review and application. *Psychol. Med.* **46,** 457–467 (2016).
14. van Loo, H. M., de Jonge, P., Romeijn, J.-W., Kessler, R. C. & Schoevers, R. A. Data-driven subtypes of major depressive disorder: a systematic review. *BMC Med.* **10,** 156 (2012).

CHAPTER 7

Casting Wider Nets for Anxiety and Depression: Disability-driven Cross-diagnostic Subtypes in a Large Cohort

Rob B. K. Wanders, Hanna M. van Loo, Jeroen K. Vermunt, Rob R. Meijer, Catharina A. Hartman, Robert A. Schoevers, Klaas J. Wardenaar, Peter de Jonge

Psychological Medicine 2016, *46:3371-3382*.

ABSTRACT

Background. In search of empirical classifications of depression and anxiety, most subtyping studies focus solely on symptoms and do so within a single disorder. This study aimed to identify and validate cross-diagnostic subtypes by simultaneously considering symptoms of depression and anxiety, and disability measures.

Methods. A large cohort of adults (LifeLines; n=73,403) had a full assessment of sixteen symptoms of mood and anxiety disorders, and measurement of physical, social and occupational disability. The best-fitting subtyping model was identified by comparing different hybrid mixture models with and without disability covariates on fit criteria in an independent test sample. The best model's classes were compared across a range of external variables.

Results. The best-fitting MM-IRT model with disability covariates identified five classes. Accounting for disability improved differentiation between people reporting isolated non-specific symptoms ('Somatic' [13.0%], and 'Worried' [14.0%]) and psychopathological symptoms ('Subclinical' [8.8%], and 'Clinical' [3.3%]). Classes showed distinct associations with clinically relevant external variables (e.g. somatization: OR=8.1-12.3, and chronic stress: OR=3.7-4.4). The 'Subclinical' class reported symptomatology at subthreshold levels while experiencing disability. No pure depression or anxiety, but only mixed classes were found.

Conclusions. An empirical classification model, incorporating both symptoms and disability identified clearly distinct cross-diagnostic subtypes, indicating that diagnostic nets should be cast wider than current phenomenology-based categorical systems.

INTRODUCTION

Despite the fact that depression and anxiety often coexist[1], rarely occur in pure form[2] and are known to load on a common internalizing dimension[3–7], the vast majority of subtyping studies have so far strictly focused on either depression[8,9] or anxiety[10,11]. Previous data-driven subtyping studies have furthermore focused solely on symptoms without taking disability into account, while disability is strongly interrelated with depression and anxiety. People with a mood or anxiety disorder experience impaired functioning on social, occupational, and physical domains[1,12–14]. The importance of disability is reflected in formal diagnostics (e.g. DSM 5) and clinical guidelines[15], where the presence of disability is a defining criterion that also signals need for care. Although increased symptom severity is reflective of the level of general functioning[16], the correlation with disability is only modest[17] and symptoms provide little domain-specific information about functional impairment (e.g. physical, social, or occupational). This is unfortunate because disability occurring in specific functional domains is known to be predictive of important clinical outcomes such as treatment response[18], remission[19,20], and recurrence[21]. This additional information conveyed in disability measures may therefore serve as a source of relevant inter-personal variation on top of the variation that is captured by symptom scores alone, and may thus have an important role in identifying subtypes.

The simultaneous consideration of a broad set of symptoms and disability within a data-driven subtyping approach could yield subtypes which better describe individual, symptom, and severity differences inherent to depression and anxiety[22–24]. Importantly, the obtained empirical classification may guide research by providing targets for biological, neurological, and genetic research[25,26], and may provide guidance as to which treatment strategy may benefit patients most[27,28].

The aim of the present study was to identify and validate data-driven cross-diagnostic subtypes to capture the heterogeneity of depressive and anxiety symptomatology in a large population sample (n=73,403), incorporating both symptoms and disability measures as sources of clinically relevant inter-personal variation. All subjects had a full assessment of their current depression and anxiety symptoms.

METHOD

PARTICIPANTS AND PROCEDURES

Data came from Lifelines[29], a large prospective population-based cohort study of 167,729 persons in the Northern Netherlands. In the study, information on a broad range of biomedical, sociodemographic, behavioral, physical and psychological factors that contribute to the health and disease of the general population are assessed, with a special focus on multimorbidity and complex genetics. Study participants were recruited via general practitioners and self-registration, following family referral to include family members. Participants visited a Lifelines research site for biomedical assessments, standardized interviews, and completed extensive questionnaires. The Lifelines cohort is broadly representative for the adult population of the north of the Netherlands, with a low risk of selection bias and good generalizability to the general population[30]. All participants signed informed consent and the Medical Ethical Committee of the University Medical Center Groningen approved the study.

For the current study, participants were included if they had complete data on an adapted version of the standardized Mini International Neuropsychiatric Interview (MINI). To facilitate quick assessments, the original MINI interview allows interviewers to skip questions about symptoms of MDD and GAD if core DSM-criteria are not met. However, between February 2012 and December 2013, Lifelines used an adapted version of the MINI in which all internalizing symptoms were assessed without symptom-skips in order to enable present analyses. This dataset consisted of 73,403 participants and was split into a training set (75%; n=55,054) for model fitting and a validation set (25%; n=18,349) for model selection and validation.

MEASURES

Internalizing symptoms

The MINI was administered by trained medical professionals and used to assess MDD, GAD, panic disorder, agoraphobia and social phobia[31]. The nine criterion-symptoms of MDD were considered present if participants experienced them almost daily during the past two weeks. The seven criterion-symptoms of GAD were considered present if experienced on most days during the past six months. Lifetime presence was assessed for agoraphobia, panic attacks and social fear. For three symptoms that overlapped between MDD and GAD ('sleep disturbance', 'fatigue' and 'concentration problems'), the GAD symptoms were excluded. The MDD symptoms were favored because of their more restrictive timeframe of two weeks, which increases the likelihood that someone who endorses one of the symptoms also currently experiences disability as a result from

that symptom. This resulted in a dataset containing 16 internalizing symptoms that were assessed for all subjects.

Disability covariates

Four subscales of the RAND-36[32] were included in the analyses: physical functioning, social functioning, role limitations due to emotional problems and role limitations due to physical health problems. The role limitations scales assess problems with work or other regular daily activities as a result of physical health or emotional problems. The RAND-36 is a widely used questionnaire to assess health-related quality of life with sound psychometric properties. Sum scores for the four scales were linearly transformed to a 0-100 range and standardized.

External variables

A wide range of external variables was assessed, including socio-demographics, lifestyle factors, psychological factors, and health status. Participants' age, gender, marital status (partner or no partner), education level (low, middle, high), income (above or below modal), and work status (currently employed, work absence, afraid to lose job; all dichotomized) were assessed using self-report questionnaires. Body Mass Index (BMI) was assessed during the research visit. Current smoking and current alcohol consumption (dichotomized as more than once a week) were assessed using self-report questionnaires. Medication use, the presence of a major medical condition (cancer, diabetes, arteriosclerosis, asthma, COPD), a cardiovascular disease (hypertension, stroke, heart failure, myocardial infarction), and cardiovascular symptoms (arrhythmia, swollen ankles, chest pain, heart function loss) were assessed using questionnaires. Somatization was assessed using a subscale of the Symptom Checklist (SCL-90[33]). Social support was measured using the subscales affection (i.e., feeling loved), behavioral confirmation (i.e., belonging and doing things right), and status (i.e., distinction in valued aspects) of the Social Production Form (SPF-IL[34]). The lifetime and past-year occurrence of stress were measured using the long-term difficulties inventory (LDI) for chronic stress and the list of threatening events (LTE) for recent stress[35]. Positive and negative affect was assessed using the positive and negative affect schedule (PANAS[36]). In line with work of Broadhead et al.[37], participants were asked to indicate in a disability days questionnaire on how many days in the preceding month they were (i) unable to perform daily activities, (ii) remained in bed, (iii) had to take a step back, or (iv) were less able to focus.

STATISTICAL ANALYSES

Missing data

Participants with more than 10 missing values on the RAND-36 (n=541) were excluded. For the remaining participants (n=73,403), missing values on RAND-36 (0.9%) and external variables (1.7%) were imputed using the R package 'mice'[38] with logistic regression for binary variables, proportional odds models for ordered categorical variables, and predictive mean matching for continuous variables (default settings).

Models

The latent structure of the included symptoms was first explored using latent class analyses (LCA) and explorative factor analyses (EFA). In LCA, a categorical latent variable representing a number of latent classes is assumed to explain all population heterogeneity in symptom-reporting without assuming severity variations within classes. In contrast to LCA, EFA is a dimensional approach which can be used to gain insight into the latent structure of the (co)variances of the assessed symptoms using continuous latent factors.

The hybrid mixture models consisted of Mixed Measurement Item Response Theory (MM-IRT) models and were fitted without disability covariates first. These models add a continuous latent factor to the LCA model accounting for quantitative differences within each class, and circumvent the problem of LCA finding parallel classes reflective of severity differences and that are not qualitatively different[9]. As such, MM-IRT provides a hybrid approach: variations in the observed data are assumed to reflect the existence of latent groups with qualitatively different symptom patterns, but within each subgroup, quantitative differences on a severity continuum are allowed. Conceptually, this means that subgroups with distinct symptom patterns exist, but that each pattern may occur at different severity levels. For example, one might find a latent class consisting of subjects who report a pattern consisting mainly out of GAD symptoms and no other symptoms. Within this group quantitative differences can then exist, with severe cases reporting all GAD symptoms and mild cases reporting only a few (e.g. only tensed but not anxious or nervous).

To incorporate information about inter-individual differences in experienced disability, both symptoms and disability were considered simultaneously in the models. For this aim, MM-IRT mixture models were run with RAND-36 measures of disability (physical functioning, social functioning, role limitations due to emotional problems, and role limitations due to physical health problems) added as covariates predicting latent class membership (Mixed Measurement Item Response Theory with Covariates; MM-IRT-C). The disability covariates were modeled in conjunction with the latent classes, and did not

serve as direct indicators of the classes, but influenced model solutions through prediction of class membership. The estimated classes still reflect different latent subgroups with distinct symptom patterns as in the MM-IRT models, but measures of disability now form an additional source of information to optimize the distinction between classes[39]. To evaluate the relative value of each disability covariate, a likelihood ratio test was used to compare the fit of the MM-IRT-C model with all disability measures to the fit of four MM-IRT-C models, each omitting one disability covariate. Next, the covariates were added in a stepwise fashion to evaluate which disability measure added most information to the model. The characteristics of the optimal model were further investigated by analyzing the associations with external variables in the full dataset.

Estimation procedure

EFA was performed using the R package 'psych'[40] and results were used to fit confirmatory factor analysis (CFA) models. LCA, CFA, MM-IRT and MM-IRT-C models were fitted in LatentGold 5.0[41]. First, models were fitted in the training set. The obtained models were then applied and compared using information criteria in the validation set, a part of the data that was not included in the training set. The use of an independent validation set for model selection prevented the selection of models that overfitted the data. All models used 500 start sets, 25000 Estimation Maximization iterations and 1000 Newton Raphson iterations until convergence in the training set. Models were run multiple times to explore likelihood values and avoid solutions at local maxima. All latent class models were nested in a single general latent variable model, allowing for direct comparisons across models. To find the optimal number of classes and select the best fitting model, the Bayesian Information Criterion (BIC) and the Consistent Akaike Information Criterion (CAIC) were compared across models in the validation set. The best LCA and EFA models served as baseline references for the hybrid mixture models. As traditional regression methods underestimate the relation between class membership and external variables and lead to biased estimates due to classification error, a corrected approach was used[42].

RESULTS

SAMPLE CHARACTERISTICS

The 73,403 participants had a mean age of 44.5 (range 18-93), 58.3% were female, and 55.1% reported at least one of the assessed symptoms (Table 1). According to the MINI, 1447 (2.0%) participants met DSM-criteria of MDD, 2983 (4.1%) met DSM-criteria of GAD, and 824 (1.1%) met DSM-criteria for both MDD and GAD. Most participants had some level of disability on at least one RAND-36 subscale (score below 100), with 36.9% reporting

disability in physical functioning below the mean (12.5% below mean minus one SD), and 26.9% experiencing social functioning below the mean (15.9% below mean minus one SD).

TABLE 1. Descriptives of baseline Lifelines sample (n=73,403).

Characteristic	Descriptive
Demographics	
Male gender, n (%)	30608 (41.7%)
Age, mean (sd)	44.5 (13.4)
Age, range	18-65
Psychiatric disorder[1]	
MDD, n (%)	1447 (2.0%)
GAD	2983 (4.1%)
Panic disorder	2455 (3.3%)
Agoraphobia	2556 (3.5%)
Social phobia	645 (0.9%)
Internalizing symptoms (MINI)	
Depressed mood, n (%)	2531 (3.4%)
Interest loss	2709 (3.7%)
Eat disturbance	3601 (4.9%)
Sleep disturbance	10292 (14.0%)
Motor disturbance	4863 (6.6%)
Fatigue	9296 (12.7%)
Guilt	1704 (2.3%)
Concentration	4607 (6.3%)
Suicidality	524 (0.7%)
Anxious	15238 (20.8%)
Nervous	9210 (12.5%)
Tense	12883 (17.6%)
Agitated	10025 (13.7%)
Panic attack	6558 (8.9%)
Agoraphobia	5398 (7.4%)
Social fear	3235 (4.4%)
Disability covariates (RAND)	
Physical functioning, mean (sd)	90.6 (14.3)
Social functioning	87.5 (18.2)
Physical role	86.7 (29.3)
Emotional role	90.7 (25.4)

SD, standard deviation; MDD, Major Depressive Disorder; GAD, Generalized Anxiety Disorder.
[1] Participants meeting criteria for current psychiatric disorder as assessed by the MINI interview.

LATENT CLASS ANALYSIS AND EXPLORATORY FACTOR ANALYSES

Comparison of LCA models with increasing numbers of classes did not point towards one definite best fitting model: information criteria kept decreasing with each class addition (Supplement 3). Inspection of the solutions with different numbers of classes showed an interesting hierarchical pattern, pointing toward the existence of subgroups with clearly different symptom prevalence rates and symptom patterns (Supplement 1). In the models with few classes, most classes were qualitatively distinct. However, with the addition of more classes, the classes that emerged showed mostly quantitative severity differences. This suggested that in these models, the latent class solutions increasingly approximated a granular representation of an underlying, continuous severity spectrum.

The EFA (Supplement 2) showed a strong first factor –a general internalizing severity factor- with a ratio between the first factor Eigenvalue (8.81) and the second factor Eigenvalue (0.82) of 10.7.

Together, the results of these explorative analyses suggested that there were qualitative as well as quantitative differences across participants in internalizing symptom reporting, indicating that the use of a hybrid MM-IRT approach could be of substantial added value.

MIXED MEASUREMENT ITEM RESPONSE THEORY MODELS

BIC and CAIC statistics indicated that MM-IRT models were superior to LCA models: the data were better described by a mixture model that allows for quantitative severity differences within classes (Supplement 3). The incorporation of disability scales as covariates further improved the model, implying that the additional information allowed the model to better distinguish between classes. Likelihood ratio tests showed that each disability covariate separately contributed significantly ($p<0.001$) to the prediction of class membership. Stepwise addition of each covariate showed that the addition of social functioning improved the model most, followed by role limitations due to emotional problems, physical functioning and role limitations due to physical health problems (see Supplement 4). Based on its lowest BIC and CAIC values in the validation set, the 5-class MM-IRT-C model was selected for further investigation. Separation between the classes in this model was strong (Supplement 5).

Variations of the final model were fitted to find out if differences in the intervals of the assessment of symptoms (last 2 weeks for MDD symptoms and last 6 months for GAD symptoms) had any influence on the model. Separate factors for symptoms assessed for the past 2 weeks and for past 6 months were included. Comparable results were found for these models and they were not preferable in terms of CAIC or BIC. Also, analyses performed using the GAD versions of the overlapping symptoms ‘sleep disturbance’, ‘fatigue’, and ‘concentration problems’, resulted in comparable findings. In addition, analyses with the

overlapping symptoms removed, retrieved the same classes except for the 'Somatic' class that was primarily based on these overlapping symptoms. Overall, this suggests that the choice of timeframe played a limited role, and that overlapping symptoms were relevant to accurate class distinction.

CLASSCHARACTERISTICS:SYMPTOMPROFILESANDEXTERNALCORRELATES

Of the 5-class MM-IRT-C model, the largest class was a 'Healthy' class of 61.0%, with a minimal amount of depression and anxiety symptoms, high physical and social functioning, scoring highest on positive psychological factors, with highest education, income and employment rate, and lowest absence from work (Figure 1, Table 2).

The model included two classes that reported mild levels of symptomatology and decreased levels of physical and social disability. The first class ('Somatic', 13.0%) reported mainly 'fatigue', and 'sleep disturbances' combined with low physical functioning. People in this class were comparatively older, had a higher BMI, poorer health status with more major medical conditions, and had few psychiatric diagnoses. The second class ('Worried', 14.0%) often reported anxiety symptoms like 'feeling anxious', and 'feeling tense' but few depression or panic symptoms. Functioning in this class was lower compared to the 'Healthy' class but considerably higher compared to the other classes and the prevalence of psychiatric disorders was low (11.1%). The 'Worried' class had increased levels of chronic and recent stress, more days with disability, more absence from work and worries about losing their job.

The two remaining classes ('Subclinical' and 'Clinical') reported higher amounts of symptomatology and higher levels of disability. The ´Subclinical´ class (8.8%) frequently reported anxiety symptoms and elevated depressive symptoms. Despite the fact that this class had poor functioning, high levels of internalizing symptoms, and low scores on positive psychological factors, only 39.5% satisfied DSM-criteria for a diagnostic classification. The 'Clinical' class (3.3%) showed high frequencies of both depression and anxiety symptoms, low levels of social functioning, and increased role limitations due to emotional problems and physical health problems. A high percentage (85.7%) of this class met criteria for a DSM-classification and almost all comorbid cases of MDD and GAD were found in this class (93.9% of the 824 comorbid cases). Interestingly, the 'Clinical' class did not capture a single disorder, but included both MDD (47.0%) and GAD (59.2%). In addition, the 'Clinical' class had poor health status compared to all but the 'Somatic' class, scored lowest on positive psychological factors, highest on chronic and recent stress, and highest on negative affect. Also, the Clinical class reported most disability days, had the lowest mean education level, income, and employment rate.

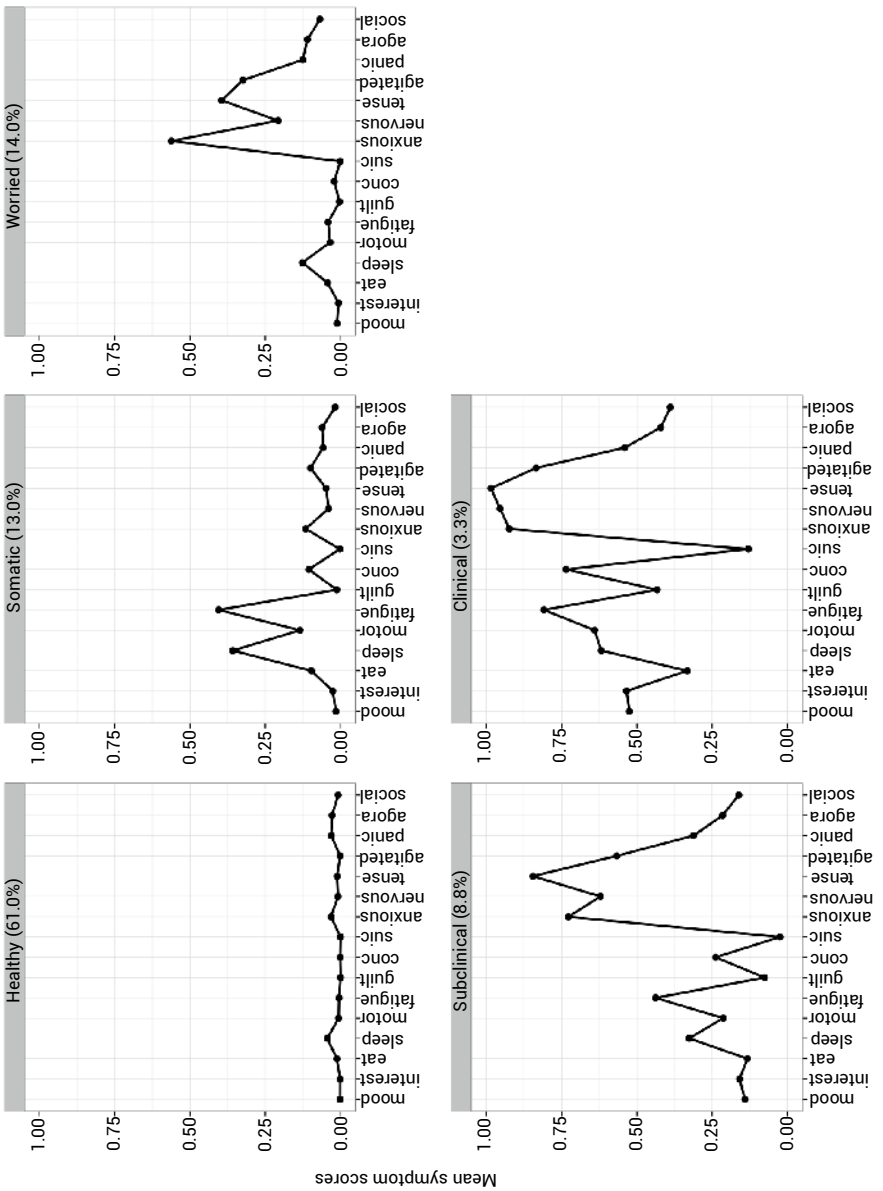


FIGURE 1. Symptom profiles of disability-driven classes of depression and anxiety symptoms of the final 5-class MM-IRT-C model.

TABLE 2. External correlates of latent classes

	Healthy	Somatic	Worried	Sub-clinical	Clinical	Cramer's V or R^2
Size	44749 (61.0%)	9519 (13.0%)	10258 (14.0%)	6438 (8.8%)	2439 (3.3%)	
Socio-demographics						
Age, mean (sd)	44.6 (13.5)	47.5 (14)	42.6 (12.7)	42.4 (12.6)	42.2 (12.1)	0.00
Male gender	47.0%	33.2%	36.8%	30.8%	26.8%	0.08
Partner	86.3%	82.7%	82.8%	78.1%	73.1%	0.02
Education - low	25.5%	36.2%	26.9%	32.1%	41.3%	0.06
Education - middle	39.5%	37.4%	41.4%	40.3%	40.8%	0.01
Education - high	33.3%	24.2%	29.9%	25.4%	15.8%	0.06
Income - below modal	37.1%	46.9%	44.7%	50.7%	60.6%	0.06
Work - Employed	82.1%	68.8%	81.8%	74.1%	59.9%	0.04
Work - Absence	7.9%	25.0%	13.8%	24.8%	39.9%	0.19
Work - Afraid to lose job	8.5%	16.2%	15.5%	22.2%	34.7%	0.15
Health status						
Medication use	40.5%	62.9%	48.4%	58.4%	67.9%	0.09
Major medical condition	33.3%	49.4%	38.2%	43.6%	48.3%	0.08
Cardiovascular disease	20.5%	31.4%	23.4%	25.6%	29.6%	0.06
Cardiovascular symptoms	38.6%	64.8%	56.4%	67.3%	79.7%	0.12
Somatization (SCL-90)	2.6 (2.7)	7.4 (5.3)	4.6 (3.8)	7.8 (5.6)	12.6 (7.6)	0.02
Lifestyle						
BMI	25.7 (4)	26.9 (5)	25.9 (4.4)	26.1 (4.8)	26.8 (5.5)	0.00
Smoking	17.9%	21.2%	23.4%	28.3%	36.6%	0.08
Alcohol (>1 week)	66.5%	55.3%	62.1%	58.1%	48.6%	0.04
Psychological						
Affection (SPF-IL)	6.6 (1.5)	6.3 (1.6)	6.2 (1.6)	5.9 (1.7)	5.3 (1.8)	0.00
Confirmation (SPF-IL)	6.4 (1.3)	6.1 (1.5)	6.1 (1.4)	5.7 (1.6)	5 (1.8)	0.00
Status (SPF-IL)	3.4 (1.6)	3.2 (1.6)	3.2 (1.6)	3 (1.7)	2.7 (1.7)	0.00
Positive affect (PANAS)	36 (3.9)	34.7 (4.4)	35.2 (4.2)	33.5 (4.8)	30.7 (5.5)	0.00
Negative affect (PANAS)	19 (4.4)	21 (4.8)	22.3 (4.7)	25.5 (5.2)	30.3 (5.6)	0.02
Acute stress (LTE)	0.88 (1.1)	1.27 (1.4)	1.32 (1.4)	1.66 (1.6)	2.24 (1.9)	0.01
Chronic stress (LDI)	1.75 (1.8)	2.88 (2.4)	3.37 (2.5)	4.55 (2.8)	6.19 (3.5)	0.03
Disability covariates (RAND)						
Physical functioning	95.1 (7.3)	75 (21.8)	92 (10.5)	85.6 (16.9)	76.3 (23.2)	0.01
Social functioning	96.2 (7.6)	74.4 (19.5)	83.1 (16.0)	68.4 (20.5)	48.1 (22.9)	0.06
Physical role	96.9 (12.8)	56.5 (41.9)	88.7 (25.4)	70.5 (39.0)	51.5 (43.7)	0.01
Emotional role	99.3 (5.2)	89.3 (25.4)	85.5 (29.1)	64.2 (40.5)	30.9 (39.1)	0.02

TABLE 2. External correlates of latent classes *(Continued)*

	Healthy	Somatic	Worried	Sub-clinical	Clinical	Cramer's V or R^2
Disability days (MINI)						
No daily activities[1]	0.3 (1.6)	1.7 (5.1)	0.5 (2.5)	1.5 (4.7)	4.7 (8.6)	0.00
Remain in bed[1]	0.2 (0.9)	0.7 (2.6)	0.3 (1.2)	0.7 (2.3)	2.2 (5.3)	0.00
Take a step back[1]	0.9 (3.4)	6 (9.3)	2.4 (5.5)	6.3 (9.1)	12.4 (11.6)	0.00
Less able to focus[1]	0.3 (1.8)	2 (5.9)	1 (3.4)	3.1 (6.6)	8.2 (10.9)	0.00
Psychiatric diagnoses (MINI)						
Any	2.2%	5.4%	11.1%	39.5%	85.7%	0.42
MDD	0.0%	0.4%	0.0%	4.1%	47.0%	0.49
GAD	0.0%	0.7%	2.4%	18.9%	59.2%	0.46
MDD & GAD	0.0%	0.1%	0.0%	0.7%	31.7%	0.47
Panic disorder	1.1%	1.8%	4.6%	11.6%	22.7%	0.23
Social phobia	0.0%	0.1%	0.7%	3.1%	14.0%	0.26
Agoraphobia	1.1%	0.1%	4.9%	10.6%	25.5%	0.24

Note. All F and X^2 tests significant at p<0.001, only effect sizes are reported.
[1] Number of days in the past 30 days.

TABLE 3. External variables predicting class-membership. Odds-ratios of standardized variables with 95% confidence intervals are reported.

	Somatic	Worried	Subclinical	Clinical	
	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	Wald sig.
Socio-demographics					
Age	1.4 (1.4-1.5)	0.9 (0.9-1.0)	1.1 (1.0-1.1)	1.1 (1.0-1.2)	<0.0001
Male gender	0.8 (0.7-0.9)	0.9 (0.8-0.9)	0.9 (0.8-1.0)	0.8 (0.7-1.0)	0.003
Partner	0.8 (0.7-1.0)	0.9 (0.9-1.1)	0.7 (0.6-0.8)	0.7 (0.6-0.8)	<0.0001
Education – low vs. high	1.3 (1.1-1.4)	1.3 (1.2-1.5)	1.5 (1.3-1.7)	2.1 (1.7-2.7)	<0.0001
Education – middle vs. high	1.0 (0.9-1.1)	1.1 (1.0-1.2)	1.1 (1.0-1.3)	1.6 (1.3-1.9)	<0.0001
Income - below modal	1.2 (1.1-1.3)	1.1 (1.0-1.2)	1.0 (0.9-1.1)	1.1 (1.0-1.3)	0.003
Work - Employed	0.7 (0.6-0.8)	1.0 (0.9-1.1)	0.7 (0.6-0.8)	0.5 (0.4-0.6)	<0.0001
Work - Absence	4.1 (3.7-4.6)	2.1 (1.9-2.4)	5.1 (4.5-5.7)	9.0 (7.6-10.6)	<0.0001
Work - Afraid to lose job	1.4 (1.2-1.5)	1.6 (1.5-1.8)	1.8 (1.6-2.1)	2.2 (1.9-2.6)	<0.0001
Health status					
Medication use	1.7 (1.6-1.9)	1.2 (1.1-1.3)	1.6 (1.4-1.8)	1.9 (1.6-2.2)	<0.0001
Major medical condition	1.3 (1.1-1.4)	1.1 (1.0-1.2)	1.2 (1.1-1.3)	1.1 (0.9-1.2)	<0.0001
Cardiovascular symptoms	1.6 (1.4-1.7)	1.6 (1.5-1.7)	1.7 (1.6-1.9)	2.3 (2.0-2.8)	<0.0001
Somatization (SCL-90)	8.5 (7.9-9.0)	3.0 (2.8-3.2)	8.1 (7.5-8.7)	12.3 (11.3-13.3)	<0.0001

TABLE 3. External variables predicting class-membership. Odds-ratios of standardized variables with 95% confidence intervals are reported. *(Continued)*

	Somatic	Worried	Subclinical	Clinical	
	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	Wald sig.
Lifestyle					
BMI	1.2 (1.1-1.2)	1.0 (1.0-1.1)	1.0 (1.0-1.1)	1.1 (1.0-1.2)	<0.0001
Smoking	1.2 (1.1-1.4)	1.3 (1.2-1.5)	1.6 (1.5-1.8)	2.1 (1.8-2.4)	<0.0001
Alcohol (>1 week)	0.7 (0.6-0.7)	0.9 (0.8-0.9)	0.7 (0.7-0.8)	0.6 (0.6-0.8)	<0.0001
Psychological					
Affection (SPF-IL)	1.0 (1.0-1.1)	1.0 (0.9-1.0)	1.0 (0.9-1.0)	0.9 (0.8-0.9)	0.002
Status (SPF-IL)	1.1 (1.0-1.1)	1.0 (1.0-1.1)	1.1 (1.1-1.2)	1.3 (1.2-1.4)	<0.0001
Positive affect (PANAS)	0.7 (0.6-0.7)	0.8 (0.8-0.8)	0.5 (0.5-0.6)	0.3 (0.3-0.4)	<0.0001
Negative affect (PANAS)	1.4 (1.3-1.5)	2.5 (2.4-2.6)	5.1 (4.7-5.4)	12.6 (11.4-13.8)	<0.0001
Acute stress (LTE)	1.1 (1.1-1.2)	1.3 (1.2-1.3)	1.3 (1.3-1.4)	1.4 (1.3-1.5)	<0.0001
Chronic stress (LDI)	2.3 (2.1-2.4)	2.8 (2.6-3.0)	3.7 (3.5-3.9)	4.4 (4.1-4.8)	<0.0001

Note: Multivariate logistic regression model with 'Healthy' class as reference estimated using proportional ML with robust standard errors. Backward selection using CAIC as criteria eliminated SPF-IL behavioral confirmation scale, and cardiovascular disease from the model (coefficient estimates were also non-significant at p>0.41).

Multivariate regression (Table 3) showed that psychological factors best predicted class membership (R^2=21.7%), followed by health status (R^2=21.3%), demographics (R^2=9.4%), and lifestyle factors (R^2=2.3%). Overall, strong predictors of 'Subclinical' and 'Clinical' class membership were work absence (OR=5.1 for the 'Subclinical' and OR=9.0 for the 'Clinical' class), somatization (OR=8.1 and 12.3), low positive affect (OR=0.3 and 0.5), high negative affect (OR=5.1 and 12.6) and chronic stress (OR=3.7 and 4.4).

DISCUSSION

Exploring symptom-data from a large population sample, the current study identified five cross-diagnostic subtypes of depression and anxiety ('Healthy', 'Somatic', 'Worried', 'Subclinical', 'Clinical'), using a statistical method that accounted for severity differences within each class and incorporated disability as an additional source of inter-personal variation. These results provided important new insights into the different phenotypical presentations of depressive and anxiety symptomatology in the population, summarized by three key observations: (i) disability is an important source of information to identify clinically relevant subtypes in psychiatry, (ii) data-driven classes did not differentiate between pure depression vs. anxiety disorders, and (iii) there was a large class with clinically relevant depressive and anxiety symptomatology at subthreshold levels. Each of these is discussed below in more detail.

First, many symptoms of depression and anxiety (e.g. fatigue) are non-specific: they frequently occur in other instances than depressive or anxiety disorders (e.g. fatigue in 'Somatic' class), which is a known source of heterogeneity in anxiety and depression[43,44]. In the current study, the inclusion of domain-specific disability measures of physical and social functioning improved the differentiation between subjects reporting isolated non-specific symptoms from those with more severe psychopathological symptoms. This finding is in line with previous studies showing the predictive value of disability measures[17], and recent ideas as stated in the RDoC framework[26] that advocate a focus broader than symptom scores alone for better distinction between normal and abnormal behavior[45].

Second, no classes describing pure depression or anxiety were observed, but only mixed presentations ('Subclinical' and 'Clinical'), consistent with growing evidence for strong etiological and phenomenological overlap between depression and anxiety. Besides strongly correlated and highly comorbid[1], both disorders share genetic risk[46], and load on a single internalizing dimension[4,47]. Moreover, depression and anxiety often have comparable treatment indications and respond similarly to antidepressants[48,49], psychosocial treatments[50] and self-guided help[51]. In contrast to previous subtyping efforts within anxiety, and within depression, the current findings suggest high convergence with etiological and treatment literatures.

Third, the finding that a large part of the sample (8.8%; 'Subclinical') experienced clinically relevant symptomatology on subthreshold levels, without meeting diagnostic criteria, is in line with studies on anxious forms of depression[52] and latent class analyses investigating mixed anxiety depression disorder[53,54]. Although labelled as subclinical in our study in line with that the majority do not meet DSM-criteria of a full diagnosis, we show that participants in this class suffer from serious disability resulting in negative impact on public health[53] and high economic costs[55,56]. Subthreshold anxiety and depression with functional impairment are moreover strong predictors of subsequent full syndrome onset[57], again emphasizing this subtype's pertinence.

The 'Subclinical' and 'Clinical' classes showed strong associations with somatization and somatic symptoms (including cardiovascular) and these associations were even stronger than in the 'Somatic' class, suggesting these somatic complaints have to some extent non-physical etiologies and are part of the mental disorder[58,59]. The 'Subclinical' and 'Clinical' classes were also characterized by markedly decreased positive affect and increased negative affect. This is in line with research showing that low positive affect and high negative affect can reflect a general vulnerability to psychopathology[56]. In the current results, this vulnerability was further emphasized by high scores on experienced chronic stress and occupational problems in these classes.

Strengths of this study include the large sample size, use of a full interview-based assessment of depression and anxiety symptoms without skips, use of advanced latent variable models, the incorporation of disability as additional source of inter-personal variation, and use of independent subsamples for model fitting (training set), and model selection and validation (validation set). However, there were also limitations. First, the use of a cross-sectional population sample without clinical information about psychiatric disorders (e.g. history, duration, treatment) limits generalizability of the results. It is likely that within the 'Clinical' class, clinically relevant subclasses also exist, but that these remained undetected because of the relatively limited number of participants with severe depression and anxiety. Second, different time intervals for current depression (last 2 weeks) and anxiety (last 6 months) were used, which might have resulted in classes which were separated based on symptom duration instead of symptom quality. However, analyses that modelled the assessment interval of each symptom found no effect on obtained results. The strength of this design is that the currently used time intervals correspond to diagnostic criteria (e.g. DSM 5) and reflect a consensus about the duration a symptom should be experienced to be considered part of a disorder.

The strong separation observed between the 'Clinical' and 'Subclinical' class suggests qualitatively distinct structures of depression and anxiety symptomatology in each class. Detailed investigation of their differences on the symptom level in future work may (i) yield insights into the specific symptoms that play a key role in the transition to a full clinical disorder with high disability[60], (ii) explain why individuals experience subthreshold symptomatology yet with associated disability, and (iii) could increase our understanding of the relation between symptoms and the underlying internalizing spectrum[7]. An interesting starting point could be symptoms of 'guilt feelings' and 'concentration problems' that showed the strongest differences between the 'Subclinical' and 'Clinical' classes. Interestingly, previous studies have found an important role for 'concentration problems' with a large unique impact across different disability domains[61,62].

The longitudinal design of Lifelines allows for future studies to test the predictive value of the identified classes in prospective data. Eventually, if these data-driven classes prove to be robust and have sufficient predictive power over time, they could serve as starting-points for new classification models and be evaluated for their potential added value for treatment. Such an alternative empirical classification scheme could serve as an early stage stratification tool, recognizing that a sizable proportion of individuals in need of care suffer from a mixed presentation of depression and anxiety at a subclinical level. Ultimately, the efficiency and effectiveness of mental health care could be improved if these 'Subclinical' individuals could be identified and treated with low-cost, effective primary care treatment programs[63] or guided self-help[51] and referring only 'Clinical' individuals to high intensity specialized care.

CONCLUSIONS

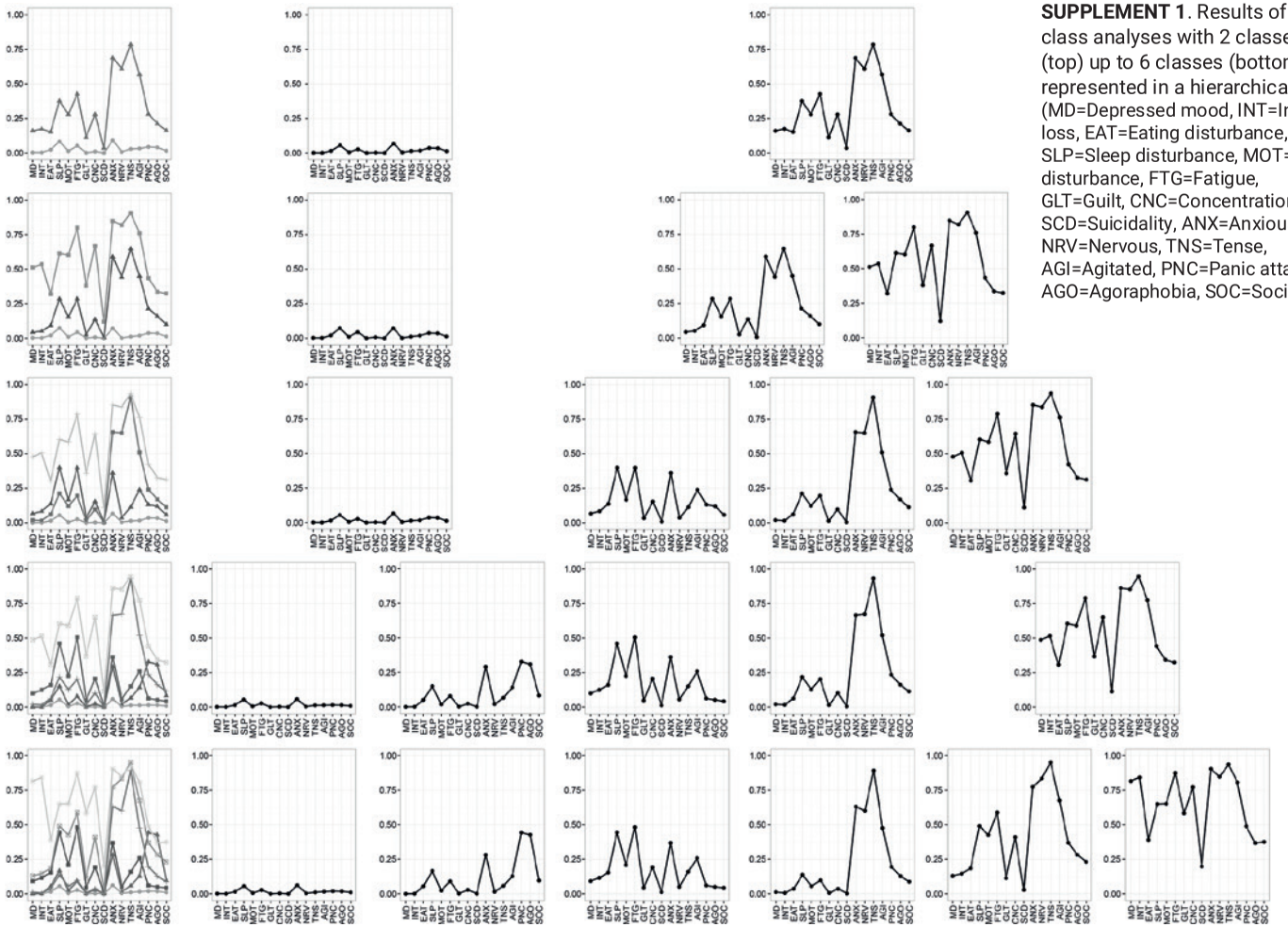
The present study described the structure of current depression and anxiety symptoms in the population with several subgroups showing mild isolated symptomatology and two cross-diagnostic subtypes with serious disability on either a clinical and subclinical level. The results suggest that diagnostic nets should be cast wider than the current phenomenology-based categorical systems to allow for a clearer focus of efforts to decrease the burden and costs of depression and anxiety in the population.

REFERENCES

1. Kessler, R. C. *et al.* Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **62,** 593–602 (2005).
2. Brown, T. A., Campbell, L. A., Lehman, C. L., Grisham, J. R. & Mancill, R. B. Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. *J. Abnorm. Psychol.* **110,** 585–599 (2001).
3. Krueger, R. F. The Structure of Common Mental Disorders. *Arch. Gen. Psychiatry* **56,** 921–926 (1999).
4. Krueger, R. F. & Bezdjian, S. Enhancing research and treatment of mental disorders with dimensional concepts: toward DSM-V and ICD-11. *World Psychiatry* **8,** 3–6 (2009).
5. Kushner, M. G. *et al.* Modeling and treating internalizing psychopathology in a clinical trial: a latent variable structural equation modeling approach. *Psychol. Med.* **43,** 1611–1623 (2013).
6. Eaton, N. R. *et al.* The structure and predictive validity of the internalizing disorders. *J. Abnorm. Psychol.* **122,** 86–92 (2013).
7. Wright, A. G. C. *et al.* The Structure of Psychopathology: Toward an Expanded Quantitative Empirical Model. *J. Abnorm. Psychol.* **122,** 281–294 (2013).
8. Sullivan, P. F., Kessler, R. C. & Kendler, K. S. Latent Class Analysis of Lifetime Depressive Symptoms in the National Comorbidity Survey. *Am. J. Psychiatry* **155,** 1398–1406 (1998).
9. van Loo, H. M., de Jonge, P., Romeijn, J.-W., Kessler, R. C. & Schoevers, R. A. Data-driven subtypes of major depressive disorder: a systematic review. *BMC Med.* **10,** 156 (2012).
10. Kessler, R. C., Stein, M. B. & Berglund, P. Social Phobia Subtypes in the National Comorbidity Survey. *Am. J. Psychiatry* **155,** 613–619 (1998).
11. Olino, T. M., Klein, D. N., Lewinsohn, P. M., Rohde, P. & Seeley, J. R. Latent trajectory classes of depressive and anxiety disorders from adolescence to adulthood: descriptions of classes and associations with risk factors. *Compr. Psychiatry* **51,** 224–235 (2010).
12. Bijl, R. V. & Ravelli, A. Current and residual functional disability associated with psychopathology: findings from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Psychol. Med.* **30,** 657–668 (2000).
13. Patten, S. *et al.* Prospective evaluation of the effect of major depression on working status in a population sample. *Can. J. Psychiatry Rev. Can. Psychiatr.* **54,** 841–845 (2009).
14. Wittchen, H. U. *et al.* The size and burden of mental disorders and other disorders of the brain in Europe 2010. *Eur. Neuropsychopharmacol.* **21,** 655–679 (2011).
15. Kramer, T., Smith, G. & Maruish, M. in *The use of psychological testing for treatment planning and outcomes assessment, 3rd ed Instruments for adults* 293–311 (2004).
16. Wakefield, J. C., Schmitz, M. F. & Baer, J. C. Does the DSM-IV Clinical Significance Criterion for Major Depression Reduce False Positives? Evidence From the National Comorbidity Survey Replication. *Am. J. Psychiatry* **167,** 298–304 (2010).
17. McKnight, P. E. & Kashdan, T. B. Purpose in life as a system that creates and sustains health and well-being: An integrative, testable theory. *Rev. Gen. Psychol.* **13,** 242–251 (2009).
18. Hirschfeld, R. M. A. *et al.* Does psychosocial functioning improve independent of depressive symptoms? a comparison of nefazodone, psychotherapy, and their combination. *Biol. Psychiatry* **51,** 123–133 (2002).
19. Von Korff, M. *et al.* Effect on disability outcomes of a depression relapse prevention program. *Psychosom. Med.* **65,** 938–943 (2003).
20. Zimmerman, M. *et al.* Remission in depressed outpatients: More than just symptom resolution? *J. Psychiatr. Res.* **42,** 797–801 (2008).
21. Scholten, W. D. *et al.* Recurrence of anxiety disorders and its predictors. *J. Affect. Disord.* **147,** 180–185 (2013).
22. Lux, V. & Kendler, K. Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria., Deconstructing major depression: a validation study of the DSM-IV symptomatic criteria. *Psychol. Med. Psychol. Med.* **40, 40,** 1679, 1679–1690 (2010).
23. Goldberg, D. The heterogeneity of 'major depression'. *World Psychiatry* **10,** 226–228 (2011).
24. Wardenaar, K. J. & de Jonge, P. Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC Med.* **11,** 201 (2013).

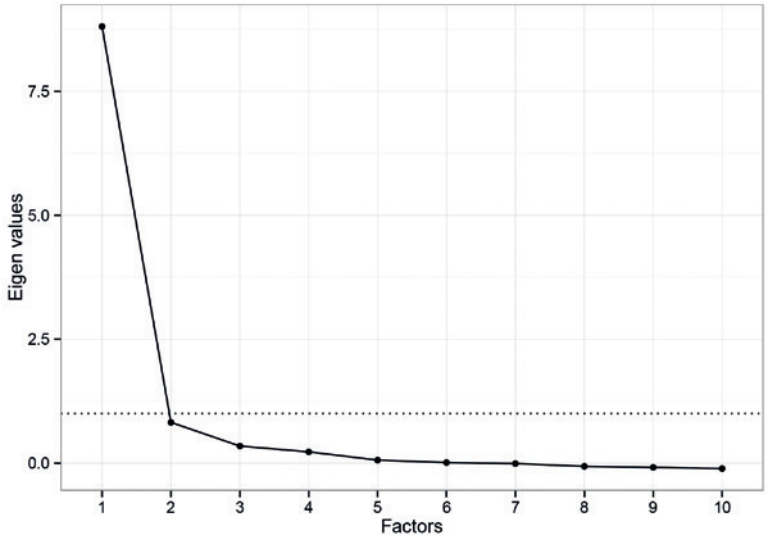
25. Kapur, S., Phillips, A. G. & Insel, T. R. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* **17,** 1174–1179 (2012).
26. Cuthbert, B. N. & Insel, T. R. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* **11,** 126 (2013).
27. Baumeister, H. & Parker, G. Meta-review of depressive subtyping models. *J. Affect. Disord.* **139,** 126–140 (2012).
28. Andrews, G., Anderson, T. m., Slade, T. & Sunderland, M. Classification of Anxiety and Depressive disorders: problems and solutions. *Depress. Anxiety* **25,** 274–281 (2008).
29. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.* **44,** 1172–1180 (2015).
30. Klijs, B. *et al.* Representativeness of the LifeLines Cohort Study. *PLOS ONE* **10,** e0137203 (2015).
31. Sheehan, D. V. *et al.* The Mini-International Neuropsychiatric Interview (M.I.N.I): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry* **59,** 22–33 (1998).
32. Hays, R. D. & Morales, L. S. The RAND-36 measure of health-related quality of life. *Ann. Med.* **33,** 350–357 (2001).
33. Derogatis, L., Lipman, R. & Covi, L. SCL-90: An outpatient psychiatric rating scale – preliminary report. *Psychopharmacol. Bull.* **9,** 13–28 (1973).
34. Nieboer, A., Lindenberg, S., Boomsma, A. & Bruggen, A. C. V. Dimensions Of Well-Being And Their Measurement: The Spf-Il Scale. *Soc. Indic. Res.* **73,** 313–353 (2005).
35. Rosmalen, J. G. M., Bos, E. H. & Jonge, P. de. Validation of the Long-term Difficulties Inventory (LDI) and the List of Threatening Experiences (LTE) as measures of stress in epidemiological population-based cohort studies. *Psychol. Med.* **42,** 2599–2608 (2012).
36. Watson, D., Clark, L. A. & Tellegen, A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Pers. Soc. Psychol.* **54,** 1063–1070 (1988).
37. Broadhead, W. E., Blazer, D. G., George, L. K. & Tse, C. K. Depression, Disability Days, and Days Lost From Work in a Prospective Epidemiologic Survey. *JAMA* **264,** 2524–2528 (1990).
38. Buuren, S. van & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **45,** (2011).
39. Tay, L., Newman, D. A. & Vermunt, J. K. Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organ. Res. Methods* **14,** 147–176 (2011).
40. Revelle, W. *psych: Procedures for personality and psychological research.* (R package version, 1(1), 2014).
41. Vermunt, J. & Magidson, J. *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax.* (Statistical Innovations Inc., 2013).
42. Vermunt, J. K. Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. *Polit. Anal.* **18,** 450–469 (2010).
43. Wanders, R. B. K. *et al.* Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make? *J. Psychosom. Res.* **78,** 130–136 (2015).
44. Wardenaar, K. J., Wanders, R. B. K., Roest, A. M., Meijer, R. R. & De Jonge, P. What does the beck depression inventory measure in myocardial infarction patients? a psychometric approach using item response theory and person-fit. *Int. J. Methods Psychiatr. Res.* **24,** 130–142 (2015).
45. Insel, T. R. The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *Am. J. Psychiatry* **171,** 395–397 (2014).
46. Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. & Eaves, L. J. Major Depression and Generalized Anxiety Disorder: Same Genes, (Partly) Different Environments? *Arch. Gen. Psychiatry* **49,** 716–722 (1992).
47. Wright, A. G. C. & Simms, L. J. A metastructural model of mental disorders and pathological personality traits. *Psychol. Med.* **45,** 2309–2319 (2015).
48. Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A. & Rosenthal, R. Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. *N. Engl. J. Med.* **358,** 252–260 (2008).
49. Roest, A. M. *et al.* Reporting Bias in Clinical Trials Investigating the Efficacy of Second-Generation Antidepressants in the Treatment of Anxiety Disorders: A Report of 2 Meta-analyses. *JAMA Psychiatry* **72,** 500–510 (2015).

50. Haby, M. M., Donnelly, M., Corry, J. & Vos, T. Cognitive behavioural therapy for depression, panic disorder and generalized anxiety disorder: a meta-regression of factors that may predict outcome. *Aust. N. Z. J. Psychiatry* **40,** 9–19 (2006).
51. Cuijpers, P., Donker, T., Straten, A. van, Li, J. & Andersson, G. Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. *Psychol. Med.* **40,** 1943–1957 (2010).
52. Goldberg, D. P. Anxious Forms of Depression. *Depress. Anxiety* **31,** 344–351 (2014).
53. Das-Munshi, J. *et al.* Public health significance of mixed anxiety and depression: beyond current classification. *Br. J. Psychiatry* **192,** 171–177 (2008).
54. Hettema, J. M., Aggen, S. H., Kubarych, T. S., Neale, M. C. & Kendler, K. S. Identification and validation of mixed anxiety–depression. *Psychol. Med.* **45,** 3075–3084 (2015).
55. Cuijpers, P. *et al.* Economic Costs of Neuroticism: A Population-Based Study. *Arch. Gen. Psychiatry* **67,** 1086–1093 (2010).
56. Ormel, J. *et al.* Neuroticism and common mental disorders: Meaning and utility of a complex relationship. *Clin. Psychol. Rev.* **33,** 686–697 (2013).
57. Karsten, J. *et al.* Psychiatric history and subthreshold symptoms as predictors of the occurrence of depressive or anxiety disorder within 2 years. *Br. J. Psychiatry* **198,** 206–212 (2011).
58. Simms, L. J., Prisciandaro, J. J., Krueger, R. F. & Goldberg, D. P. The structure of depression, anxiety and somatic symptoms in primary care. *Psychol. Med.* **42,** 15–28 (2012).
59. Bekhuis, E., Boschloo, L., Rosmalen, J. G. M. & Schoevers, R. A. Differential associations of specific depressive and anxiety disorders with somatic symptoms. *J. Psychosom. Res.* **78,** 116–122 (2015).
60. Fried, E. I. & Nesse, R. M. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med.* **13,** 72 (2015).
61. Tweed, D. L. Depression-related impairment: estimating concurrent and lingering effects. *Psychol. Med.* **23,** 373–386 (1993).
62. Fried, E. I. & Nesse, R. M. The Impact of Individual Depressive Symptoms on Impairment of Psychosocial Functioning. *PLOS ONE* **9,** e90311 (2014).
63. Patel, V. *et al.* Effectiveness of an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in Goa, India (MANAS): a cluster randomised controlled trial. *The Lancet* **376,** 2086–2095 (2010).



SUPPLEMENT 1. Results of latent class analyses with 2 classes (top) up to 6 classes (bottom) represented in a hierarchical view. (MD=Depressed mood, INT=Interest loss, EAT=Eating disturbance, SLP=Sleep disturbance, MOT=Motor disturbance, FTG=Fatigue, GLT=Guilt, CNC=Concentration, SCD=Suicidality, ANX=Anxious, NRV=Nervous, TNS=Tense, AGI=Agitated, PNC=Panic attack, AGO=Agoraphobia, SOC=Social fear)

SUPPLEMENT 2. Screenplot for the Exploratory Factor Analyses of all depression and anxiety symptoms, with the dashed line representing an eigenvalue of 1. Eigenvalues for the first four factors were 8.81, 0.82, 0.34 and 0.23 respectively, with a ratio of first to second factor of 10.7. In the 1-factor model, all symptoms showed strong factor loadings above 0.3 with an average loading of 0.49 (range 0.32-0.74). In the 2-factor model, not all symptoms showed loadings above 0.3, fit indices did not reach criteria of good fit (CFI=0.89, TLI=0.86), and factors were highly intercorrelated (r=0.72). Taken together, this points at a strong general factor where the dimensionality of the data can be sufficiently described by a single continuous factor.



SUPPLEMENT 3. Comparison of fitted models.

	Log Likelihood	Npar	ΔBIC[1]	ΔCAIC[1]	Smallest class
LCA					
2-class	-196252	33	9118	9082	0.20
3-class	-190887	50	5902	5883	0.05
4-class	-188327	67	4537	4534	0.05
5-class	-187539	84	4309	4324	0.04
6-class	-186899	101	3876	3908	0.02
7-class	**-186406**	**118**	**3855**	**3904**	**0.02**
FA[2]					
1-factor	-190577	33	5570	5534	
2-factors	-187359	33	3648	3612	
MM-IRT					
2-class	-187128	65	3828	3824	0.12
3-class	**-186011**	**98**	**3497**	**3526**	**0.07**
4-class	-185557	131	3644	3706	0.09
5-class	-185288	164	3820	3917	0.06
6-class	-185160	197	4042	4170	0.02
7-class	-185054	203	4276	4437	0.02
MM-IRT-C Disability[2]					
2-class	-181505	69	0	0	0.22
3-class	-178338	106	-1638	-1601	0.06
4-class	-177155	143	-2067	-1993	0.05
5-class	**-176287**	**180**	**-2249**	**-2138**	**0.03**
6-class	-175759	217	-2111	-1963	0.03
7-class	-175509	254	-1967	-1782	0.02
(Reference)			(120892)	(120961)	

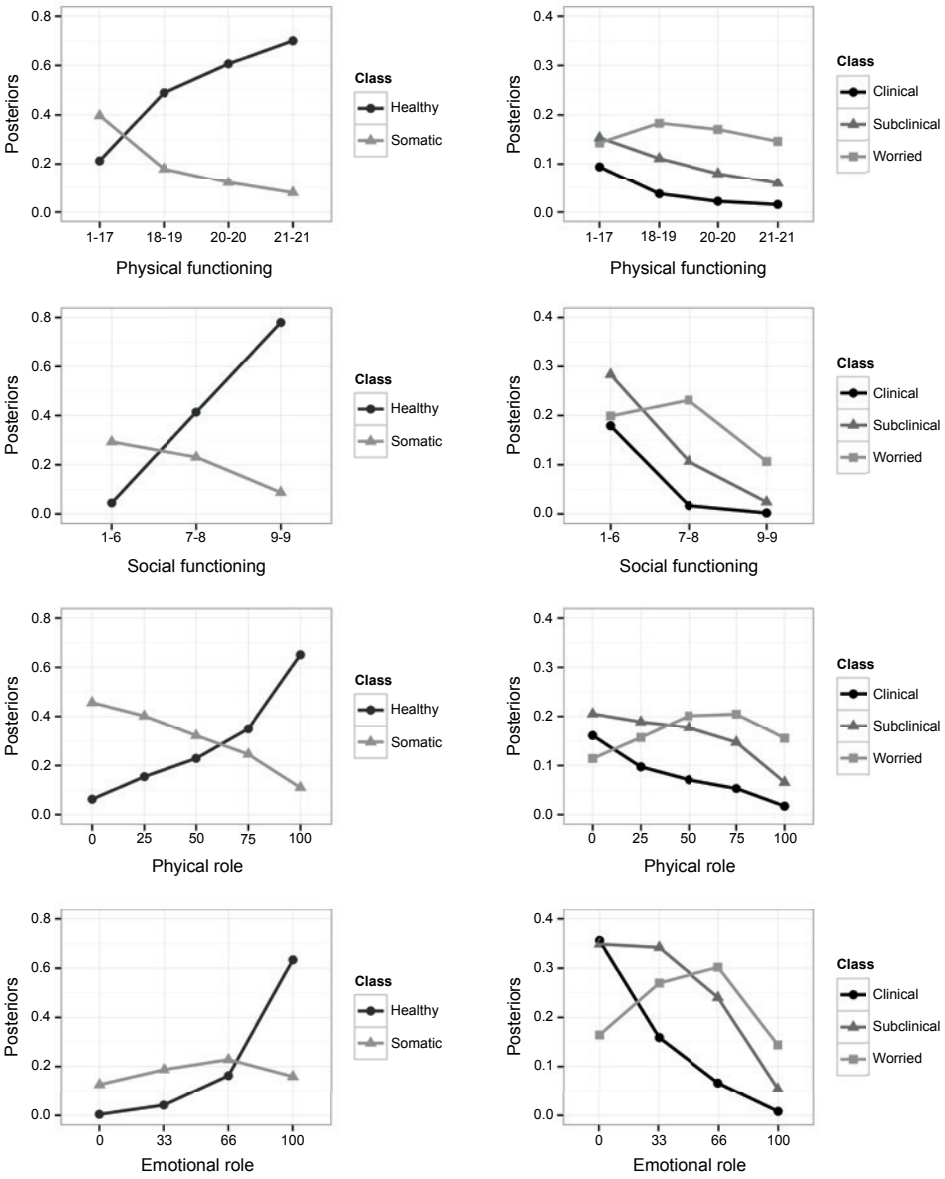
BIC, Bayesian Information Criterion; CAIC, Corrected Akaike Information Criteria; Npar, number of parameters; LCA, Latent Class Analysis; FA, Factor Analysis; MM-IRT, Mixed Measurement Item Response Theory;

[1] Differences in BIC and CAIC with respect to the '2-class MM-IRT-C Disability' model are reported to allow for easier model comparison (i.e. positive difference indicate worse fit and vice versa). The interpretation of BIC and CAIC remains the same, since only relative differences in information criteria are meaningful, and still allows direct comparisons.

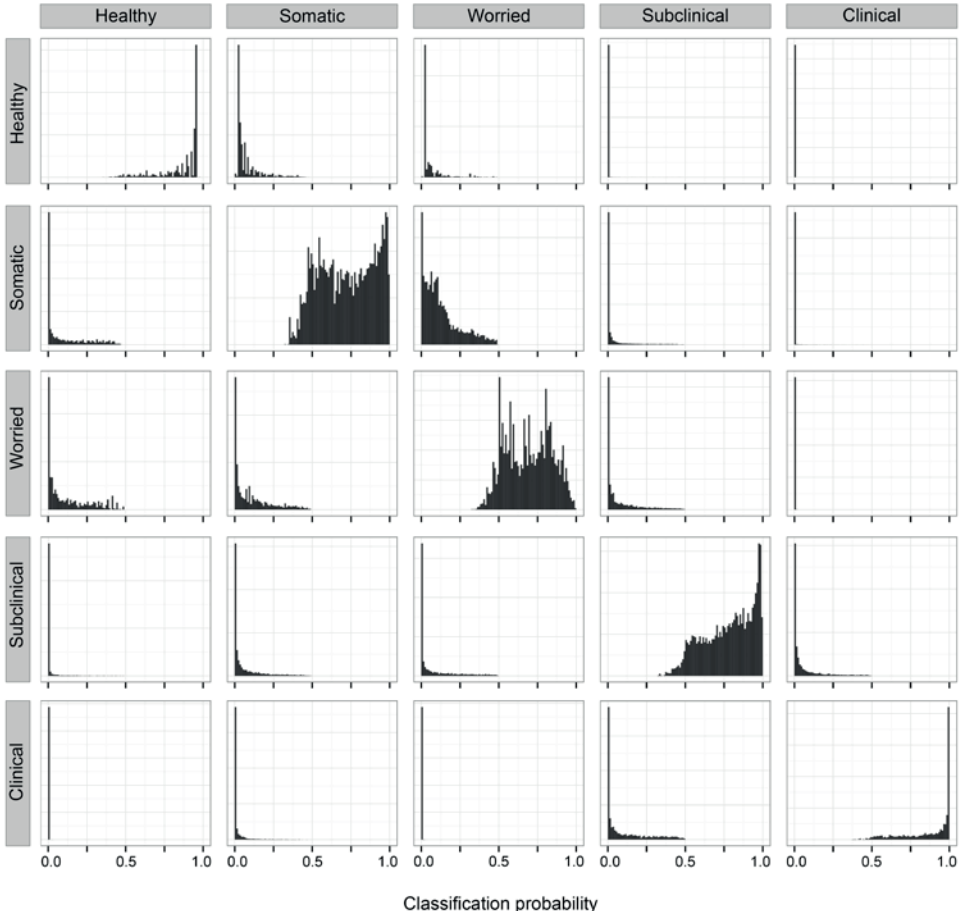
[2] Confirmatory factor analyses models based on EFA results with Promax rotation. Note that information criteria are a poor means to compare factor analysis models, and we refer to supplement 2 for more details.

[2] Disability covariates of RAND-36 subscales physical functioning, social functioning, emotional role and physical role limitations.

SUPPLEMENT 4. Posterior probabilities of class membership in the final 5-class MM-IRT-C model, associated with different levels on each disability covariate (row wise), with scales ranging from poor (0) to good (100) functioning. Left plots show probabilities for 'Healthy' and 'Somatic' class, and right for 'Worried', 'Subclinical', and 'Clinical' class. These probabilities show the role disability plays in assigning participants to each specific class. Good functioning on all scales is associated with a high chance of being assigned to the 'Healthy' class, poor physical functioning with a high chance of ending up in the 'Somatic' class. More subtle roles of disability are observed in the remaining three classes. Interestingly, participants with severe role limitations due to emotional problems have the highest probability of getting assigned to the 'Clinical' class, despite being the smallest class.



SUPPLEMENT 5. For each class all posterior probabilities of getting assigned to that class and each other class are plotted. Classes are well separated if subjects assigned to a class have high probabilities to be assigned in that particular class, and low probabilities to be assigned to the remaining other classes. The plot shows that classes are well separated (Entropy of 0.79), with especially the 'Subclinical' and 'Clinical' classes highly discriminative. The median posterior probability was 0.93 for 'Healthy' class, 0.73 for 'Somatic', 0.70 for 'Worried', 0.81 for 'Subclinical', and 0.95 for 'Clinical'.



CHAPTER 8

Patterns and Dimensionality of Depressive and Anxiety Symptomatology in the General Population

Klaas J. Wardenaar, Rob B.K. Wanders,
Margreet ten Have, Ron de Graaf,
Peter de Jonge

Under review

ABSTRACT

Background. Previously, researchers have tried to identify more homogeneous subtypes of major depressive disorder (MDD) with latent class analyses (LCA). However, this approach does no justice to the dimensional nature of psychopathology. In addition, anxiety and functioning-levels have seldom been integrated in subtyping efforts. Therefore, this study used a hybrid discrete-dimensional approach to identify subgroups with shared patterns of depressive and anxiety symptomatology, while accounting for functioning-levels.

Method. The Comprehensive International Diagnostic Interview (CIDI) 1.1 was used to assess previous-year depressive and anxiety symptoms in the Netherlands Mental Health Survey and Incidence Study-1 (NEMESIS-1). The symptom data (n=5583) were analyzed with factor analyses, LCA and hybrid mixed-measurement item response theory (MM-IRT) models with and without functioning covariates. Finally, predictors (measured one year earlier) and 2-year outcomes of class-membership were investigated.

Results. A 3-class MM-IRT model with functioning-scales as covariates best described the data. This model consisted of a healthy class (74.2%), and two symptomatic classes: a sleep/energy class (13.4%) and a mood/anhedonia class (12.4%). Factors including age, urbanicity, severity and 1-year MDD predicted being in either of the symptomatic classes rather than the healthy class one year later. In addition, both symptomatic classes generally showed poorer 2-year outcomes (i.e. disorders, poor functioning) than the healthy class. More specifically, the mood/anhedonia class showed higher odds of MDD after two years than the sleep/energy class.

Conclusion. Heterogeneity in depression and anxiety symptomatology can be optimally described by a hybrid discrete-dimensional subtyping model. Accounting for functioning-levels further helps to capture relevant interpersonal differences.

INTRODUCTION

Depression is a common and burdensome condition for which the underlying etiological mechanisms and optimal treatment are still poorly understood. An important reason for the lack of scientific progress so far may lie in the fact that the traditionally used depression diagnosis is very heterogeneous and unsuitable to capture all relevant inter-individual variation among those suffering from depressive symptoms[1,2].

To overcome the problem of diagnostic heterogeneity, empirical studies have used statistical models such as Latent Class Analysis (LCA) to identify data-driven subtypes of depression that are characterized by different symptom-probability patterns. These studies have shown that depressed patients can be subdivided into more homogeneous subtypes[3–14]. Moreover, several studies found that the identified classes showed different patterns of association with, for instance, treatment response[14], biomarkers[15] and depression course[16]. However, the interpretability of LCA results is hampered by the underlying key-assumption that all heterogeneity is explained by a finite number of classes and no additional (co)variation exist within classes. Due to this, continuous variations and/or symptom-clustering/covariance within subgroups are not captured with LCA models. This leads to rather crude subtyping models that do little justice to the real-life observation that psychopathology is in fact a largely continuous phenomenon[3,17].

To better account for the dimensional nature of psychopathology when identifying subtypes, an alternative, hybrid mixture approach could be used. One such hybrid approach is mixed measurement item response theory (MM-IRT [18–20]), which integrates LCA with an underlying IRT measurement model. MM-IRT is closely related to factor mixture models (FMM [21–23]), which have been used to integrate LCA with a factor-analytic approach[24–29]. In MM-IRT, a measurement (IRT) model is taken as point of departure and heterogeneity in response behavior is explained by estimating latent classes for which different IRT-model parameters may hold. More specifically, the same IRT model is assumed to hold for the whole population, but latent classes may exist with different estimated parameter values[30,31]. Covariates can be included in the MM-IRT models as well to further improve the differentiation between identified classes (MM-IRT-C [32]). In the hybrid MM-IRT approach, LCA and IRT complement each other. On the one hand, latent population heterogeneity can be investigated while accounting for between- and within-class variations in response-behavior. On the other hand, the underlying dimensionality of the symptoms can be investigated, while accounting for latent population heterogeneity (Clark et al., 2013). MM-IRT has been used in a variety of settings; for instance, to investigate heterogeneity in response behavior on personality questionnaires[33–35], patterns of tobacco-use/dependence symptoms[22] and the use of special response scales[36].

Apart from the fact that they have disregarded underlying dimensionality, another issue of many previous subtyping studies is that they have only analyzed symptoms of DSM-defined MDD. However, depressive symptoms are known to seldom occur in isolation, but often co-occur with other 'internalizing' symptoms of anxiety[37]. Therefore, a cross-diagnostic approach focusing on depressive and anxiety symptomatology together in a single analysis is likely to provide a more complete insight into inter-individual variations in symptomatology. Indeed, results from a recent subtyping study indicated that anxiety symptoms may play an important role in the differentiation between depression subtypes[29]. Another important aspect that has gained relatively little attention in the subtyping literature is that individuals can differ strongly in terms of the functional limitations that they experience with their symptoms. Although functioning levels and/or disability are considered very important to determine whether symptoms are actually indicative of a pathological state[38,39], their role as a source of heterogeneity on top of variations in depressive/anxiety symptom patterns has hardly been investigated in data-driven subtyping studies. This is unfortunate because accounting for functioning levels could, for instance, help to better distinguish between groups of people with 'clinical' and 'sub-clinical' levels of disability.

In a recent subtyping study that used data from a large cohort (n=73,403), Wanders et al. [40] addressed the above-described issues by using MM-IRT to estimate subgroups based on depressive and anxiety symptoms, while accounting for the role of functioning levels. The results showed that a 5-class MM-IRT-C model with functioning scales incorporated as covariates, optimally described heterogeneity and differentiated between subgroups with different depression and anxiety symptom profiles ('healthy', 'somatic', 'worried', 'subclinical' and 'clinical' subgroups). Interestingly, the identified classes also showed different patterns of cross-sectional associations with external factors, such as sociodemographic, lifestyle and psychological factors. These results clearly showed the added value of using the hybrid MM-IRT approach in data-driven subtyping work.

The current study aimed to use a similar approach, applying MM-IRT-C to depressive and anxiety symptomatology to identify cross-diagnostic subtypes, while incorporating measures of functioning as covariates. However, this study also aimed to extend on the previous work by making use of a representative population sample (n=5583; the Netherlands Mental Health Survey and Incidence Study-1, NEMESIS-1). In addition, the 3-wave longitudinal design of NEMESIS-I (baseline, 1-year follow-up and 3-year follow-up) allowed for a thorough investigation of the longitudinal correlates of the estimated classes. MM-IRT-C models were estimated on the data collected at 1-year follow-up. Next, the prediction of subgroup membership by variables measured earlier at baseline was

investigated. Finally, the ability of the identified subgroups to predict outcomes measured at 3-year follow-up was investigated.

METHODS

PARTICIPANTS

Participants came from NEMESIS-1, a longitudinal cohort study in a randomly selected adult population sample (aged 18-65 years) from the Netherlands. The study consisted of a baseline measurement (T0; n=7,076; 69.7% response; in the year 1996) a measurement after 1 year (T1; n=6,518; 79% response; in the year 1997) and a measurement after 3 years (T2; n=4,796; 85% response; in the year 1999). The detailed design, rationale and goals of NEMESIS-1 have been described previously[41]. The research protocol was approved by the central medical ethics committee for mental health (METIGG). All participants provided oral informed consent in line with the prevailing Dutch law at the time the fieldwork took place.

In each measurement-wave, participants were interviewed with the (Composite International Diagnostic Interview, CIDI; version 1.1) generating DSM-III-R diagnoses. The depression questions (Section E) and anxiety symptom questions (Section D: Panic Disorder, Generalized Anxiety Disorder [GAD], Agoraphobia, Social Phobia and Specific phobia) were used in the current study. The T1 data (collected 1 year after the baseline measurement) were used to estimate an optimal subtyping model because the time frame of these CIDI symptom assessments was limited to the 1-year period between T0 and T1 (see Figure 1). This ensured that the assessed symptoms (co)occurred roughly within the same 1-year time-interval. A previous multivariate analysis showed that sample attrition between T0 and T1 was associated with younger age, lower education, urbanicity, not cohabiting with a steady partner, unemployment, being born outside the Netherlands, agoraphobia, social phobia and eating disorders. The presence of any DSM-III-R disorder was only weakly related to attrition, controlled for demographics (OR=1.20; [42]). Of the 5,618 respondents at T1, 5583 (99.4%) provided all the data that was needed for the current analyses (measures of depressive and anxiety symptomatology and functioning). The current analyses were conducted using data from the different measurement waves (see Figure 1). The MM-IRT and MM-IRT-C analyses were run in the T1 sample, identifying a range of subgroups. Next, factors assessed at T0 were used to predict subgroup-membership at T1. Finally, the subgroups at T1 were associated with outcomes measured at T2 to evaluate the prognostic value of the identified subgrouping. This was done using all subjects that were included in the MM-IRT analyses and had T2 assessments for the

relevant outcomes (see below). When adjusted for sociodemographic factors, attrition between T1 and T2 was associated with the presence of MDD, dysthymia and alcohol dependence[43].

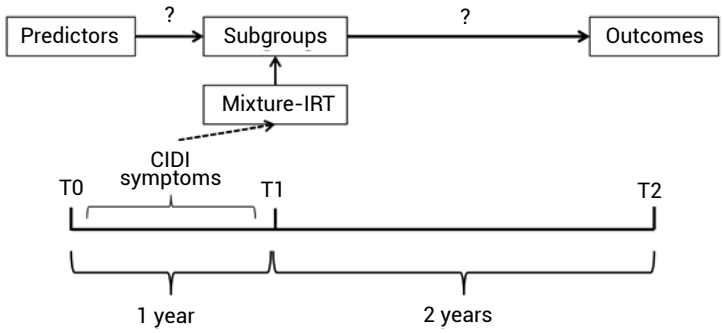


FIGURE 1. Design of the study: First, MM-IRT is used to identify subgroups based on symptoms assessed at T1 for the period T0-T1 with the CIDI. Next, T0 predictors of the T1 subgroups are investigated. Finally, the T1 subgroups' predictions of T2 outcomes are investigated.

MEASUREMENTS

Symptom-assessments and functioning at T1

The presence of depressive symptoms was assessed with the depression section of the CIDI 1.1. All depressive symptoms were evaluated irrespective of whether the key-symptoms were endorsed (there were no symptom skips in the depression section). In addition, the responses to CIDI screening questions for a range of common anxiety disorders (Panic Disorder, GAD, Agoraphobia, Social Phobia and Specific phobia) were also used in the current study. Here, only the screening questions could be used because the anxiety sections of the CIDI skipped all detailed questions if the screening questions were not endorsed. Taken together, the analyzed symptom-dataset contained 28 depressive symptoms and 5 anxiety symptoms.

The Medical Outcome Study Short Form-36 (SF-36 [44]) was used to assess several domains of functioning.

Predictors at T0

Sociodemographic variables were assessed at baseline and included: age, gender, employment status and educational attainment. The Mastery scale[45] was used to assess the extent to which individuals feel in control and/or feel responsibility for the events occurring in their lives (internal vs. external 'locus of control'), with higher scores indicating more external locus of control. The Rosenberg Self Esteem scale[46] was used to assess self-esteem. The General Health Questionnaire-12 (GHQ-12[47]) was used to

assess severity. In addition, 1-year CIDI-based DSM-III-R diagnoses were determined, using disorder hierarchies and exclusion rules. In addition, the SF-36 was used to assess functioning.

Outcomes at T2

Five health-related outcomes were constructed based on assessments at T2. The CIDI was used to confirm the presence of a 1-year MDD diagnosis and any 1-year anxiety diagnosis at T2. In addition, functioning outcomes were based on the SF-36. Poor psychological functioning was defined as scoring in the lowest tertile of the SF-36 psychological health scale and poor social functioning was defined as scoring in the lowest tertile of the SF-36 social functioning scale. Poor physical functioning was defined as scoring in the lowest tertile of the SF-36 physical functioning scale.

STATISTICAL ANALYSES

A step-by-step approach was taken to identify the data-driven model that best described the symptom-data heterogeneity. First, EFA and LCA were conducted to explore the latent structure of the data. Next, MM-IRT models were run. The MM-IRT-C models were run using the following scales of the SF-36 as covariates: physical functioning, social functioning, physical role functioning and emotional role functioning. These four scales were selected out of the eight SF-36 scales for use as covariates because they were expected to add unique relevant information about the subjects' health-related functioning, independently of the information about the subjects' mental states that was provided by the symptom data. Both role-functioning scales were dichotomized (100%=1 and <100%=0).

Factor analyses were run using Mplus (v 7.0 [48]) and mixture analyses were conducted with Latent GOLD (v.5.0 [49]). Multiple random starts were used in the estimation of the mixture models to avoid identification of models at local maxima.

Exploratory Factor Analyses and Latent Class Analyses

Weighted least squares Exploratory Factor Analysis (EFA) was run on the tetrachoric symptom correlation matrix. The ratios between the first and subsequent factors' eigenvalues were inspected to evaluate if the data were best described by a one factor model or by more factors.

To investigate the optimal number of classes to describe the heterogeneity in the current sample, LCA was conducted. Models with increasing numbers of classes were estimated and the Bayesian Information criterion (BIC) and Akaike Information Criterion (AIC) were compared to identify the model that best described the data.

Mixture IRT

MM-IRT models were first fitted to the data without covariates. In these models, a continuous latent dimension (or more than one latent dimension if the EFA showed a multifactorial structure) was modeled in addition to the basic LCA model to account for quantitative severity differences within each class. In addition, the measurement-model parameters were allowed to vary across classes. The resulting model allowed for the identification of latent classes with qualitatively different patterns of symptom endorsement but also allowed for quantitative severity variations within each class. The latter means that within each class symptom-patterns can occur at different severity levels. After estimation of the regular MM-IRT models, several MM-IRT-C models were estimated with sets of SF-36 functioning scales added as covariates. First, the social and physical functioning scales were added. Second, the physical and emotional role-functioning scales were added. Finally, all four scales were added as covariates.

External associations

After identification of the model that best described the data, its associations with external variables were investigated. These analyses consisted of two steps. In the first step, associations between variables at T0 and class membership at T1 were investigated. Univariate and multivariate multinomial regression analyses were run, using the categorical latent class variable as outcome and sociodemographic, psychiatric and psychological variables at T0 as independent variables. In the second step, class membership at T1 was used to predict the health-related outcomes at T2. Logistic regression models were run with each of these outcomes as dependent variable and the latent class dummy-variables as independent variables. Crude models were run first, unadjusted for any covariates. Next, models were rerun with demographic variables assessed at T1 (i.e. gender, age, education, urbanicity, living situation and employment) added as covariates. Finally, the models were rerun with additional covariates assessed at T1 to adjust for severity (GHQ-12 score) and each target outcome's initial level at T1.

Weighting

All analyses (EFA, LCA, MM-IRT and external associations) were run using the appropriate post-stratification weights[41].

RESULTS

BASELINE DESCRIPTIVE INFORMATION

Of the selected 5583 participants, 49.7% was female and the mean age was 39.3 years. Of the sample, 64.6% was employed, 15.7% was homemaker, 7.1% was student, 6.2% was unemployed or disabled and 6.3% was retired. Of the participants, 82.5% percent came from an urban area, 69.1% lived together with a spouse and 30.1% had a college education. In the year prior to baseline, 5.6% had MDD, 2.3% had panic disorder, 1.0% had GAD, 1.2% had agoraphobia without panic disorder, 4.4% had social phobia, and 6.9% had specific phobia.

FACTOR ANALYSIS

In the EFA, Eigenvalues of the first five factors were 18.4, 1.7, 1.6, 1.3 and 1.1, respectively. The ratio between the Eigenvalues of the first and second extracted factor was 11.5, indicating that the first factor by far described most of the variance in the data. Inspection of the factor loadings in the 1-factor model showed that all items had substantial factor loadings (range: 0.44-0.89). Based on these results, a single latent dimension was modeled in the subsequent MM-IRT models.

Latent Class Analysis

The LCA results are shown in Table 1. The BIC decreased with each class addition, up to 6 classes, whereas the AIC decreased further with more class additions. Based on these results, the 6-class model was selected for further inspection (Supplement 1). The model consisted of a 'healthy' (69.7), 'sleep problems' (6.7%), 'lack of energy' (13.7%), 'moderate somatic depression' (5.1%), 'moderate cognitive depression' (2.8%) and 'severe' (2.0%) class.

Mixture IRT

The results of the mixture IRT analyses are shown in Table 1. The MM-IRT models had lower BIC and AIC values than the LCA models, indicating that the hybrid approach of MM-IRT led to a better description of the data. Also, MM-IRT models with two or more classes had lower BIC/AIC values than the single-class MM-IRT model, indicating that modeling of multiple classes with different IRT parameters led to a better description of the data than a single IRT-model for the whole sample.

Lower BIC and AIC values indicated that the MM-IRT-C models described the data better than the regular MM-IRT models. Of these MM-IRT-C models, the 3-class model with all four functioning scales added as covariates described the data best, as shown by the lowest BIC value. The model entropy was 0.76. This model was selected for further investigation. The classes were characterized by different symptom-endorsement patterns (see Figure 2). The first and largest class was characterized by low endorsement of all

symptoms (healthy: 74.2%). The second class (sleep/energy: 13.4%) was characterized by comparatively high endorsement of sleeping problems, morning tiredness and energy loss. The third class (mood/anhedonia: 12.4%) also showed increased endorsement of sleep problems and energy loss, but in addition, showed high endorsement rates of loss of (sexual) interest, anhedonia and problems with concentration/decision-making.

TABLE 1. Results of fitting latent class models and MM-IRT models to the depression and anxiety symptom data at T1.

Model	Covariates	Model	Log likelihood	Number of parameters	BIC	AIC
LCA	-	2-class	-33564.8	67	67707.6	67263.5
		3-class	-31980.4	101	64832.3	64162.9
		4-class	-31614.3	135	64393.4	63498.6
		5-class	-31382.6	169	64223.4	63103.2
		6-class	-31223.0	203	64197.4	62851.9
		7-class	-31089.5	237	64223.9	62653.1
MM-IRT	-	1-class	-31946.6	66	64462.7	64025.3
		2-class	-31231.8	133	63651.0	62769.5
		3-class	-30925.9	200	63577.4	62243.7
		4-class	-30720.7	267	63745.0	61975.6
MM-IRT-C	PF & SF	2-class	-30943.1	135	63050.9	62156.1
		3-class	-30608.4	204	62977.0	61624.9
		4-class	-30352.7	273	63060.8	61251.4
	PRF & ERF	2-class	-30951.2	135	63067.2	62172.5
		3-class	-30626.8	204	63013.7	61661.6
		4-class	-30392.9	273	63141.3	61331.8
	PF, SF, PRF & ERF	2-class	-30844.2	137	62870.5	61962.5
		3-class	-30512.8	208	62820.3	61441.7
		4-class	-30270.3	279	62947.8	61098.6

All analyses were weighted. LCA=Latent class analysis; MM-IRT=Mixed Measurement Item response theory; MM-IRT-C=MM-IRT with covariate(s); BIC=Bayesian Information Criterion; AIC=Akaike Information Criterion; PF=SF36 physical functioning scale; SF=SF-36 social functioning scale; PRF=physical role functioning; ERF=emotional role functioning.

Each of the four covariates showed a significant association with class membership: social functioning (Wald statistic=179.6; p<0.001), physical functioning (Wald statistic=10.4; p=0.006), physical role functioning (Wald statistic=8.0; p=0.02), and emotional role functioning (Wald statistic=185.9; p<0.001). Inspection of the posterior probabilities of class-membership for different levels of functioning (Supplement 2) showed that scoring high on each of the scales was associated with a higher probability of being classified in the healthy class. Lower scores were associated with being classified in one of the two symptomatic classes.

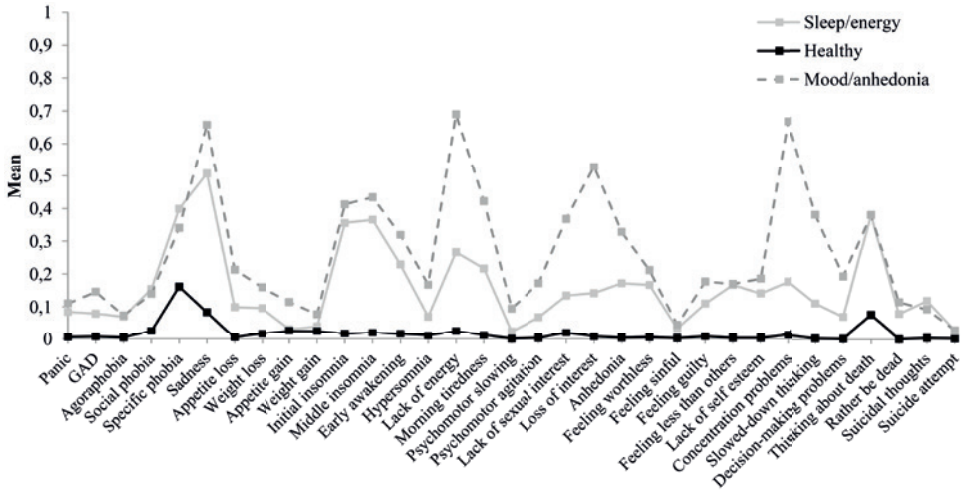


FIGURE 2. Mean item scores in each of the estimated classes of the 3-class MM-IRT-C model

CLASS-MEMBERSHIP PREDICTION

Multinomial regression was performed to investigate the predictive associations of a range of variables that were measured at T0 with class-membership at T1. In these analyses, both symptomatic classes were first compared to the healthy class (Table 2) and next compared to each other (Table 3). Univariate analyses showed that almost all investigated predictors were significantly associated with increased odds of being in the sleep/energy class or mood/anhedonia class compared to the healthy class. In multivariate analyses, several of these effects were no longer significant. Higher odds of being in the mood/anhedonia class rather than the healthy class at T1 were associated with female gender (OR=1.48), urban environment (OR=1.35), having paid employment (OR=1.34), higher GHQ-12 score (OR=1.11), a 1-year MDD diagnosis (OR=3.12), 1-year panic disorder (OR=1.85), 1-year GAD (OR=2.11), 1-year agoraphobia (OR=2.04), 1-year specific phobia (OR=1.65), lower SF-36 social functioning (OR=0.92), lower SF-36 vitality (OR=0.91), lower SF-36 psychological health (OR=0.81), lower SF-36 general health (OR=0.93) and lower mastery (OR=0.96) at T0. Higher odds of being in the sleep/energy class rather than the healthy class at T1 were associated with female gender (OR=1.52), urban environment (OR=1.46), living alone (OR=1.38), higher GHQ-score (OR=1.08), 1-year MDD (OR=3.19), 1-year panic disorder (OR=2.14), 1-year agoraphobia (OR=2.26), 1-year specific phobia (OR=1.46), lower SF-36 psychological health (OR=0.87), lower SF-36 general health (OR=0.93) and lower self-esteem (OR=0.95) at T0.

TABLE 2. Prediction of class membership at T1 by clinical and psychological predictors measured one year earlier at T0.

	Class 1: Healthy (n=4236)		Class 2: Mood/Anhedonia (n=656)			Class 3: Sleep/energy (n=694)		
				Univariate	Multivariate		Univariate	Multivariate
	Descriptives		Descriptives	OR (95%CI)	OR (95%CI)	Descriptives	OR (95%CI)	OR (95%CI)
Female gender, n (%)	1924 (45.4%)	ref.	393 (59.9%)	1.80 (1.52-2.13)***	1.48 (1.22-1.79)***	429 (61.8%)	1.95 (1.65-2.29)***	1.52 (1.26-1.83)***
Age, mean (s.d.)[b]	39.0 (12.7)	ref.	39.0 (11.3)	1.00 (0.97-1.04)	1.00 (0.96-1.04)	40.5 (12.6)	1.05 (1.02-1.09)**	1.03 (0.99-1.07)
Urban environment, n (%)	3448 (81.4%)	ref.	568 (86.7%)	149 (1.17-1.89)**	1.35 (1.04-1.74)*	605 (87.2%)	1.55 (1.22-1.96)***	1.46 (1.13-1.88)**
College education, n (%)	1255 (29.6%)	ref.	200 (30.5%)	1.04 (0.87-1.24)	-	212 (30.5%)	1.05 (0.88-1.24)	-
Living alone, n (%)	666 (15.7%)	ref.	143 (21.8%)	1.49 (1.22-1.83)***	1.11 (0.87-1.40)	167 (24.1%)	1.70 (1.40-2.06)***	1.38 (1.11-1.72)**
Paid employment, n (%)	2785 (65.7%)	ref.	409 (62.3%)	0.86 (0.73-1.02)	1.34 (1.09-1.64)**	381 (54.9%)	0.64 (0.54-0.75)***	1.04 (0.86-1.26)
GHQ-12 score, median (IQR)	0.0 (0.0-1.0)	ref.	1.0 (0.0-4.0)	1.41 (1.37-1.46)***	1.11 (1.06-1.17)***	1.0 (0.0-3.0)	1.33 (1.28-1.38)***	1.08 (1.03-1.14)**
1-year MDD, n (%)	85 (2.0%)	ref.	115 (17.5%)	10.35 (7.71-13.9)***	3.12 (2.20-4.41)***	110 (15.8%)	9.17 (6.82-12.33)***	3.19 (2.26-4.50)***
1-year Panic Disorder, n (%)	34 (0.8%)	ref.	44 (6.8%)	9.09 (5.76-14.34)***	1.85 (1.05-3.27)*	48 (6.9%)	9.22 (5.88-14.44)***	2.14 (1.23-3.71)**
1-year GAD, n (%)	18 (0.4%)	ref.	22 (3.4%)	8.21 (4.39-15.34)***	2.11 (1.01-4.42)*	15 (2.1%)	5.08 (2.55-10.15)***	1.42 (0.64-3.16)
1-year Agoraphobia, n (%)	23 (0.5%)	ref.	21 (3.2%)	6.17 (3.40-11.20)***	2.04 (1.02-4.07)*	24 (3.4%)	6.49 (3.63-11.60)***	2.26 (1.17-4.38)*
1-year Social phobia, n (%)	103 (2.4%)	ref.	66 (10.0%)	4.47 (3.24-6.16)***	1.21 (0.80-1.82)	76 (10.9%)	4.93 (3.62-6.71)***	1.45 (0.99-2.14)
1-year Specific phobia, n (%)	189 (4.5%)	ref.	101 (15.4%)	3.90 (3.02-5.05)***	1.65 (1.20-2.27)**	93 (13.4%)	3.31 (2.55-4.31)***	1.46 (1.06-2.01)*
Functioning scales, median (IQR)								
SF-36 Physical functioning[a]	100 (95-100)	ref.	95 (85-100)	0.81 (0.78-0.85)***	1.01 (0.94-1.08)	95 (85-100)	0.78 (0.75-0.82)***	0.96 (0.90-1.03)
SF-36 Social functioning[a]	100 (90-100)	ref.	87.5 (67.5-100)	0.68 (0.65-0.71)***	0.92 (0.86-0.97)**	90 (67.5-100)	0.72 (0.69-0.75)***	0.97 (0.91-1.03)
SF-36 Vitality[a]	75 (65-85)	ref.	60 (45-75)	0.64 (0.61-0.67)***	0.91 (0.85-0.98)*	65 (50-80)	0.68 (0.65-0.71)***	0.95 (0.89-1.02)
SF-36 Pain[a]	100 (80-100)	ref.	90 (67.5-100)	0.83 (0.81-0.86)***	0.98 (0.93-1.03)	90 (67.5-100)	0.82 (0.80-0.85)***	0.96 (0.92-1.01)
SF-36 Psychological health[a]	88 (80-92)	ref.	76 (60-84)	0.54 (0.51-0.57)***	0.81 (0.74-0.88)***	76 (60-88)	0.59 (0.56-0.62)***	0.87 (0.80-0.95)**
SF-36 General health[a]	80 (70-90)	ref.	70 (55-85)	0.73 (0.70-0.76)***	0.93 (0.87-0.99)*	70 (55-85)	0.73 (0.70-0.76)***	0.93 (0.87-0.98)*
Mastery scale, mean (s.d.)	20.0 (3.0)	ref.	17.8 (3.9)	0.83 (0.81-0.85)***	0.96 (0.93-1.00)*	17.9 (3.7)	0.83 (0.81-0.85)***	0.97 (0.94-1.01)
Rosenberg Self-esteem scale, mean (s.d.)	33.6 (3.8)	ref.	31.2 (4.8)	0.87 (0.85-0.88)***	0.99 (0.97-1.02)	31.0 (4.7)	0.86 (0.84-0.87)***	0.95 (0.93-0.98)***

OR=odds ratio; 95%CI= 95% Confidence Interval; GHQ-12=General Health Questionnaire-12; MDD= Major Depressive Disorder; GAD=Generalized Anxiety Disorder, MOS-SF 36=Medical Outcome Study Short Form-36.

[a] Odds ratios are given for 10-point increments.

[b] Odds ratios given for 5-year increments

*p<0.05, **p<0.01; ***p<0.001

TABLE 3. Prediction of membership of the 'sleep/energy' vs. 'mood/anhedonia' classes.

	T1 class-membership		
	Sleep/energy (n=694)	Mood/Anhedonia (n=656)	
		Univariate[a]	Multivariate[a]
T0 predictors		**OR (95%CI)**	**OR (95%CI)**
Female gender, n (%)	Ref	0.93 (0.74-1.15)	-
Age, mean (s.d.)	Ref	0.95 (0.91-0.99)*	0.95 (0.91-1.00)*
Urban environment, n (%)	Ref	0.96 (0.70-1.32)	-
College education, n (%)	Ref	1.05 (0.94-1.17)	-
Living alone, n (%)	Ref	0.88 (0.68-1.13)	-
Paid employment, n (%)	Ref	1.36 (1.09-1.69)**	1.40 (1.12-1.76)**
GHQ-12 score, median (IQR)	Ref	1.07 (1.03-1.11)***	1.04 (0.99-1.10)
1-year MDD, n (%)	Ref	1.32 (0.96-1.83)	-
1-year Panic Disorder, n (%)	Ref	1.01 (0.65-1.58)	-
1-year GAD, n (%)	Ref	1.62 (0.80-3.28)	-
1-year Agoraphobia, n (%)	Ref	0.82 (0.43-1.56)	-
1-year Social phobia, n (%)	Ref	0.97 (0.67-1.40)	-
1-year Specific phobia, n (%)	Ref	1.19 (0.87-1.62)	-
Functioning scales, median (IQR)			
SF-36 Physical functioning	Ref	1.04 (0.98-1.10)	-
SF-36 Social functioning	Ref	0.94 (0.90-0.99)**	0.97 (0.91-1.03)
SF-36 Vitality	Ref	0.93 (0.89-0.98)*	0.99 (0.92-1.08)
SF-36 Pain	Ref	1.01 (0.97-1.06)	-
SF-36 Psychological health	Ref	0.91 (0.86-0.97)**	0.96 (0.87-1.05)
SF-36 General health	Ref	1.00 (0.94-1.05)	-
Mastery scale, mean (s.d.)	Ref	0.99 (0.96-1.02)	-
Rosenberg Self-esteem scale, mean (s.d.)	Ref	1.01 (0.99-1.03)	-

GHQ-12=General Health Questionnaire-12; MDD= Major Depressive Disorder; GAD=Generalized Anxiety Disorder, SF-36=Medical Outcome Study Short Form-36; IQR=interquartile range.

[a] Logistic regression analyses using class-membership (0=sleep/energy and 1=mood/anhedonia) as outcome.

[b] OR given per 5-year increase.

[c] OR given per 10-point increase.

*p<0.05; **p<0.01; ***p<0.001

TABLE 4. Predictive associations between class-membership at T1 with clinical outcomes two years later at T2

	T2 Outcomes									
	1-year MDD diagnosis		1-year Anxiety diagnosis		Lowest tertile: SF-36 Psychological health		Lowest tertile: SF-36 Social functioning		Lowest tertile: SF-36 Physical functioning	
Prediction models	**n**	**OR (95% CI)**	**n**	**OR (95% CI)**	**n**	**OR (95%CI)**	**n**	**OR (95%CI)**	**n**	**OR (95%CI)**
Crude										
Mood/anhedonia vs. rest	4768	7.32 (5.34-10.04)***	4768	5.38 (3.99-7.27)***	4768	3.33 (2.77-4.00)***	4768	2.88 (2.39-3.47)***	4768	2.03 (1.69-2.44)***
Sleep/energy vs. rest	4768	3.54 (2.45-5.13)***	4768	5.25 (6.90-7.08)***	4768	3.04 (2.54-3.64)***	4768	2.77 (2.30-3.32)***	4768	2.19 (1.83-2.62)***
Mood/anhedonia vs. sleep/energy	1161	2.07 (1.41-3.03)***	1161	1.03 (0.74-1.43)	1161	1.09 (0.87-1.38)	1161	1.04 (0.82-1.32)	1161	0.93 (0.73-1.18)
Adjusted 1										
Mood/anhedonia vs. rest	4708	6.85 (4.97-9.45)***	4708	5.10 (3.74-6.95)***	4708	3.12 (2.58-3.76)***	4708	2.67 (2.21-3.23)***	4708	2.01 (1.65-2.44)***
Sleep/energy vs. rest	4708	3.15 (2.15-4.63)***	4708	5.06 (3.71-6.90)***	4708	2.74 (2.28-3.30)***	4708	2.42 (2.00-2.92)***	4708	1.96 (1.62-2.38)***
Mood/anhedonia vs. sleep/energy	1144	2.23 (1.50-3.30)***	1144	1.03 (0.73-1.45)	1144	1.15 (0.90-1.46)	1144	1.11 (0.87-1.41)	1144	1.02 (0.80-1.31)
Adjusted 2										
Mood/anhedonia vs. rest	4702	4.30 (2.89-6.40)***	4702	2.37 (1.63-3.47)***	4698	1.19 (0.94-1.49)	4702	1.47 (1.18-1.84)**	4702	1.54 (1.22-1.94)***
Sleep/energy vs. rest	4702	2.65 (1.78-3.95)***	4702	2.80 (1.98-3.96)***	4698	1.41 (1.14-1.74)**	4702	1.67 (1.36-2.05)***	4702	1.43 (1.15-1.78)**
Mood/anhedonia vs. sleep/energy	1144	1.79 (1.17-2.72)**	1141	0.89 (0.61-1.29)	1139	0.92 (0.71-1.19)	1141	0.95 (0.73-1.23)	1141	1.09 (0.82-1.44)

OR=Odds ratio; 95%CI=95% Confidence Interval. SF-36=Medical Outcome Study Short Form-36.
The results are based on multinomial regression analyses using the healthy class as reference category.
Adjusted 1: T1 gender, age, urbanicity, living alone, employment status
Adjusted 2: T1 gender, age, urbanicity, living alone, employment status, GHQ-12 score and status on the outcome at T1

Univariate comparisons of the symptomatic classes using 'sleep/energy' as reference (Table 3), showed that a lower age (OR=0.95), having paid employment (OR=1.36), higher GHQ-12 scores (OR=1.07), lower social functioning scores (OR=0.94), lower vitality scores (OR=0.93) and lower psychological health (OR=0.91) at T0 were all significantly associated with a higher odds of being in the mood/anhedonia class than in the sleep/energy class at T1. In the multivariate analyses only a lower age (OR=0.95) and having paid employment (OR=1.40) at T0 retained significant predictive effects.

PREDICTION OF T2 OUTCOMES

Predictive associations of class membership at T1 with outcomes two years later at T2 are shown in Table 4. In the unadjusted models, membership of the 'sleep/energy' class and the 'mood/anhedonia' class were both associated with higher odds of a MDD diagnosis, any anxiety diagnosis, poor psychological health, poor social functioning and poor physical functioning at T2. The OR's decreased but mostly remained significant when prediction models were adjusted for demographic covariates at T1 and for covariates to adjust for the outcome's status at T1.

Logistic regression analyses to compare the odds of the outcomes between the two symptomatic classes showed that only the odds of a MDD diagnosis at T2 were higher in the 'mood/anhedonia' class than in the 'sleep/energy' class (OR=1.79). Together, these results indicate that membership of any symptomatic class was predictive of adverse outcomes but that the mood/anhedonia and sleep/energy classes only differed in terms of their risk of depression at follow-up.

DISCUSSION

This study aimed to find an optimal subtyping model to describe heterogeneity in depressive and anxiety symptomatology in a population sample. The results showed that a 3-class hybrid discrete-dimensional MM-IRT-C model with functioning scales added as covariates best described the data. The resulting model divided the sample into a large healthy class, a sleep/energy class and a mood/anhedonia class. In two steps, associations of the classes with external variables were investigated. In the first step, analyses to investigate if class-membership at T1 could be predicted using assessments from a year earlier (T0) showed that factors including female gender, urban environment, living alone, GHQ-12 severity, the presence of 1-year MDD, the presence of several 1-year anxiety disorders (panic disorder, agoraphobia and specific phobia), SF-36 psychological health and SF-36 general health were all predictive of being in either of the symptomatic

classes vs. the healthy class a year later. More specifically, a 1-year GAD diagnosis, lower SF-36 social functioning, SF-36 vitality and mastery scores at T0 predicted higher odds of being in the mood/anhedonia class rather than the healthy class. A lower self-esteem score was specifically predictive of being in the sleep/energy class rather than the healthy class. However, prospective differentiation between the symptomatic classes appeared hard: only age and employment status at T0 were found to differ between the sleep/energy or mood/anhedonia classes. In the second step, longitudinal analyses to investigate the predictive value of class membership for outcomes two years later showed that both being in the sleep/energy class and the mood/anhedonia class was associated with poorer outcome compared to the healthy class. Comparisons of the symptomatic classes showed that being in the mood/anhedonia class was predictive of higher odds of 1-year MDD at follow-up compared to the sleep/energy class. On the other outcomes, no such differentiation was between symptomatic classes was observed. Taken together, these results indicated that MM-IRT with covariates identified population subgroups that were clearly different in terms of their symptom-endorsement patterns. Importantly, predictors and outcomes of class-membership were shown to differ between the healthy and the two symptomatic classes, and between the two symptomatic classes (mood/anhedonia vs. sleep/energy). Several of these results' implications are discussed below.

The fact that a MM-IRT model fit the data better than a LCA model showed the importance of accounting for the dimensionality of depression and anxiety symptomatology when aiming to optimally subtype persons based on their symptom-reporting. The finding that a MM-IRT model with 3 classes fit the data better than the best-fitting LCA model with 6 classes is in line with previous notions that discrete-dimensional hybrid models often provide more parsimonious solutions (with fewer classes) than purely discrete LCA models, especially when the symptoms have a strong underlying (uni)dimensional structure[50]. The findings also align with the previous MM-IRT study[40], although that study identified a MM-IRT model with more classes, including a purely somatic class, a subclinical class and more specific symptomatic classes. The current findings can also be compared with those from other studies using a closely related FMM approach. Sunderland et al.[27] found that in a large sample of treatment-seeking patients (n=1165), hybrid FMMs better described heterogeneity in depressive symptom-data than regular LCA models, as indicated by lower AIC and BIC values, although the authors concluded that a more parsimonious factor model outperformed the FMMs. Conversely, in a symptomatic population sample (NEMESIS-II; n=1388), ten Have et al.[29] found that a LCA with a freely estimated within-class correlation between appetite gain and appetite-loss, described depression data better than a range of tested hybrid FMMs. Variations in the findings across the referenced studies as well as the current study could be related to differences in sample composition (more healthy vs.

more symptomatic; select vs. random samples), and thus, in the frequency distributions of the input-variables for the statistical models. Wanders et al.[40], for instance, found five rather than three classes with MM-IRT-C in a sample that was recruited via general practitioners. This selection is likely to have resulted in a sample with much more reported symptomatology than the current randomly selected population sample. The resulting higher absolute numbers of symptomatic cases require that more classes are added to the model to optimally explain all the relevant symptom-pattern heterogeneity than when using data from a population sample with low symptom-rates. Another aspect that could explain differences between studies' results are the time-interval for which symptoms were assessed varied across studies (past two weeks[27] to worst lifetime episode[29]). Also, differences in the used assessment instruments (Patient Health Questionnaire [PHQ-9][27] vs. CIDI[29] vs. Mini International Neuropsychiatric Interview [MINI][40]) and the exact approach of the analyses could play a role. For instance, ten Have et al.[29] used a regular LCA and an LCA-variant that allows for some local dependence within the model, whereas Sunderland et al.[27] and the current study only employed regular LCA.

Another interesting finding in the current study was that the model provided a better description of the data when the SF-36 (role-)functioning scales were added as covariates to the model. This aligns with the previous study by Wanders et al.[40], and shows that functioning/disability levels should be considered an important source of heterogeneity, in line with existing ideas about the relevance of disability as an indicator of pathological problems and need for care[38,39]. Secondly, these findings form a proof-of-principle for the notion that the incorporation of more available information in a subtyping model helps to improve the data-driven explanation of heterogeneity in a sample. Remarkably, this has received little attention in the depression subtyping literature. Previous studies have focused on the identification of subtypes that are purely based on symptom-reports. The resulting subtypes' associations with other variables are only investigated in post-hoc analyses to evaluate the (external) validity of the symptom-based subtypes[4,11,29]. The current results show that it might be preferable to already include other relevant information (i.e. functioning, medication-use, demographics etc.) in the actual subtyping models. In that way, more sources of inter-individual variation are incorporated in- and explained by the model, leading to subtypes that are not only better in terms of statistical fit, but also in terms of differentiation between different clinical pictures. Although mixture models can become very complex when multiple sources of variation are added to it, including the additional variables as covariates in the current study appeared like a suitable way to incorporate more information into the model at the expense of only a few added model parameters.

In addition to the abovementioned methodological implications, the current finding of two distinct symptomatic classes, 'sleep/energy' vs. 'mood/anhedonia', is also interesting from a theoretical perspective. The distinction between persons with more pronounced mood-related and cognitive symptoms and persons with mostly somatic and/or vegetative symptoms has been observed previously in data-driven studies of depression heterogeneity. Several symptom-based factor-analytical studies have shown that depression symptomatology can be decomposed into a mood/cognition-related symptom factor and a somatic symptom factor[51,52]. Interestingly, these domains were shown to be differently related with prospective outcomes, with mood-cognitive symptomatology being specifically predictive of the presence of MDD and somatic symptoms being specifically predictive of anxiety at 2-year follow-up[53]. This aligns with the current finding that being in the mood/anhedonia class was specifically predictive of MDD at follow-up. However, anxiety at follow-up did not differ significantly between the classes, possibly due to the fact that both classes showed similar levels of sleep-problems and relatively little difference in endorsed anxiety symptoms. Previous person-based studies have also found evidence for the distinct roles of mood/cognitive and somatic symptoms in the course-heterogeneity of depression. For instance, Monden et al.[54] reported that by using multiway principle component analysis, longitudinal depression data could be decomposed into different person-components: one with persisting cognitive/affective symptoms, one with persisting somatic symptoms and one with quick recovery on both domains. Another study showed that depressive patients could be subdivided into data-driven classes that were characterized by different combinations of course-trajectories on somatic vs. mood/cognitive symptomatology[55]. Although different in terms of the used models and study-populations, all of the abovementioned work seems to indicate that a considerable part of the heterogeneity in depression is related to the distinction between mood/cognitive and somatic/vegetative symptoms.

The results showed no clear distinction between subgroups reporting only depressive symptoms and subgroups reporting only anxiety symptoms: both symptomatic classes showed roughly similar levels of anxiety. This finding is in line with the results by Wanders et al.[40], who also found no specific depressive or anxiety subgroups. Also, these findings fit in the larger literature on depression and anxiety comorbidity, showing that the disorder groups more often occur together than alone and very likely share common underlying risk-factors and etiological mechanisms[37].

Interestingly, unlike previous LCA studies that have found atypical and typical subtypes, characterized by appetite/weight-gain and appetite/weight-loss, respectively[11], the current classes did not show such a strong distinction. This is likely to be due to the different ways

in which local dependence between appetite/weight gain and loss (with a correlation close to -1) influences LCA and hybrid models. The problem of local independence in LCA can lead to the identification of artificial classes, neutralizing local dependence by appointing those with appetite/weight-loss to one class and those with appetite/weight-gain to another class. This issue was recently illustrated by ten Have et al.[29], who showed that accounting for the local dependence between appetite/weight loss and appetite/weight gain in a LCA model led to a model that was statistically better, but no longer had classes that reflected differences in appetite/weight loss and gain. In MM-IRT and FMM, (co) variation within classes is allowed and modeled with an IRT or factor model. This might weaken the local dependencies and prevent the differentiation between classes based on locally dependent symptoms alone. As such, these hybrid models can be expected to yield parsimonious models with fewer classes than LCA-based models.

Although this study had several strengths, including the substantial sample size, CIDI data without any skipped depression symptoms, the availability of predictors that were measured a year earlier and the availability of outcomes that were measured two years later, there were also some study limitations that should be kept in mind. First, the results apply to part of a population sample and generalizability to other (clinical) samples needs to be further investigated. Second, the data came from a relatively old survey using an older version of the CIDI and DSM. However, this limitation was outweighed by the fact that full CIDI depressive symptom-assessments were available for all subjects because no symptom-skips were used. In addition, the current study focused on the latent structure of symptomatology, which is unlikely to have changed since the data was collected. Third, the time frame in which the assessed symptoms could occur was one year, which meant that reported symptoms did not necessarily all occur at the same time. However, by using the T1 data, it was at least made sure that the interval between reported symptoms never exceeded one year. Fourth, the SF-36 scales that were added as covariates were assessed at T1, whereas the assessed symptoms could have occurred anywhere between T0 and T1. Fifth, the sample-size was not large enough to allow for a random split into a model-development sample and a model-validation sample for confirmatory modeling. In future studies, the results of the current study should therefore be replicated in independently collected data. In addition, the explored models could be extended with a broader range of internalizing and externalizing psychopathological symptoms and by inclusion of more or other covariates.

In conclusion, the heterogeneity in depression and anxiety symptomatology was best described by a hybrid discrete-dimensional subtyping model. Moreover, incorporating information about persons' functioning was shown to lead to a subtyping model that even

better described interpersonal heterogeneity. Apart from these findings' implications for depression and anxiety subtyping, the current study should mainly be seen as a proof-of-principle for (1) the use of MM-IRT in psychopathology subtyping and (2) the use of clinical covariates in data-driven subtyping models. Ultimately, the use of these flexible analytical approaches could contribute to the identification of psychopathology subtypes that combine adequate fit to empirical data with optimal usefulness in research and clinical settings.

REFERENCES

1. Widiger, T. A. & Clark, L. A. Toward DSM-V and the classification of psychopathology. *Psychol. Bull.* **126,** 946–963 (2000).
2. Wardenaar, K. J. & de Jonge, P. Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC Med.* **11,** 201 (2013).
3. Kendell, R. Clinical validity. *Psychol. Med.* **19,** 45 (1989).
4. Kendler, K. S. *et al*. The Identification and Validation of Distinct Depressive Syndromes in a Population-Based Sample of Female Twins. *Arch. Gen. Psychiatry* **53,** 391–399 (1996).
5. Sullivan, P. F. & Kendler, K. S. Typology of common psychiatric syndromes. An empirical study. *Br. J. Psychiatry* **173,** 312–319 (1998).
6. Sullivan, P. F., Kessler, R. C. & Kendler, K. S. Latent Class Analysis of Lifetime Depressive Symptoms in the National Comorbidity Survey. *Am. J. Psychiatry* **155,** 1398–1406 (1998).
7. Sullivan, P. F., Prescott, C. A. & Kendler, K. S. The subtypes of major depression in a twin registry. *J. Affect. Disord.* **68,** 273–284 (2002).
8. Parker, G., Wilhelm, K., Mitchell, P., Roy, K. & Hadzi-Pavlovic, D. Subtyping depression: testing algorithms and identification of a tiered model. *J. Nerv. Ment. Disord.* **187,** 610–7 (1999).
9. Carragher, N., Adamson, G., Bunting, B. & McCann, S. Subtypes of depression in a nationally representative sample. *J. Affect. Disord.* **113,** 88–99 (2009).
10. Hybels, C. F., Blazer, D. G., Pieper, C. F., Landerman, L. R. & Steffens, D. C. Profiles of Depressive Symptoms in Older Adults Diagnosed With Major Depression: Latent Cluster Analysis. *Am. J. Geriatr. Psychiatry* **17,** 387–396 (2009).
11. Lamers, F. *et al.* Identifying Depressive Subtypes in a Large Cohort Study: Results From the Netherlands Study of Depression and Anxiety (NESDA). *J. Clin. Psychiatry* **71,** 1582–1589 (2010).
12. Lamers, F. *et al.* Structure of major depressive disorder in adolescents and adults in the US general population. *Br. J. Psychiatry* bjp.bp.111.098079 (2012). doi:10.1192/bjp.bp.111.098079
13. Li, Y. *et al.* Subtypes of major depression: latent class analysis in depressed Han Chinese women. *Psychol. Med.* **44,** 3275–3288 (2014).
14. Ulbricht, C. M., Rothschild, A. J. & Lapane, K. L. The association between latent depression subtypes and remission after treatment with citalopram: A latent class analysis with distal outcome. *J. Affect. Disord.* **188,** 270–277 (2015).
15. Lamers, F. *et al.* Evidence for a differential role of HPA-axis function, inflammation and metabolic syndrome in melancholic versus atypical depression. *Mol. Psychiatry* **18,** 692–699 (2013).
16. Lamers, F., Beekman, A. T. F., Hemert, A. M. van, Schoevers, R. A. & Penninx, B. W. J. H. Six-year longitudinal course and outcomes of subtypes of depression. *Br. J. Psychiatry* **208,** 62–68 (2016).
17. Kendell, R. & Jablensky, A. Distinguishing Between the Validity and Utility of Psychiatric Diagnoses. *Am. J. Psychiatry* **160,** 4–12 (2003).
18. Rost, J. Rasch models in latent classes: An integration of two approaches to item analysis. *Appl. Psychol. Meas.* **14,** 271–282 (1990).
19. Rost, J. A logistic mixture distribution model for polychotomous item responses. *Br. J. Math. Stat. Psychol.* **44,** 75–92 (1991).
20. Mislevy, R. J. & Verhelst, N. Modeling item responses when different subjects employ different solution strategies. *Psychometrika* **55,** 195–215 (1990).
21. Lubke, G. H. & Muthén, B. Investigating Population Heterogeneity With Factor Mixture Models. *Psychol. Methods* **10,** 21–39 (2005).
22. Muthen, B. & Asparouhov, T. Item response mixture modeling: Application to tobacco dependence criteria. *Addict. Behav.* **31,** 1050–1066 (2006).
23. Lubke, G. H. & Miller, P. J. Does nature have joints worth carving? A discussion of taxometrics, model-based clustering and latent variable mixture modeling. *Psychol. Med.* **45,** 705–715 (2015).
24. Lubke, G. H. *et al.* Subtypes Versus Severity Differences in Attention-Deficit/Hyperactivity Disorder in the Northern Finnish Birth Cohort. *J. Am. Acad. Child Adolesc. Psychiatry* **46,** 1584–1593 (2007).

25. Kuo, P.-H., Aggen, S. H., Prescott, C. A., Kendler, K. S. & Neale, M. C. Using a factor mixture modeling approach in alcohol dependence in a general population sample. *Drug Alcohol Depend.* **98,** 105–114 (2008).

26. Picardi, A. *et al.* Heterogeneity and symptom structure of schizophrenia. *Psychiatry Res.* **198,** 386–394 (2012).

27. Sunderland, M., Carragher, N., Wong, N. & Andrews, G. Factor mixture analysis of DSM-IV symptoms of major depression in a treatment seeking clinical population. *Compr. Psychiatry* **54,** 474–483 (2013).

28. Pattyn, T. *et al.* Identifying Panic Disorder Subtypes Using Factor Mixture Modeling. *Depress. Anxiety* **32,** 509–517 (2015).

29. ten Have, M. *et al.* The identification of symptom-based subtypes of depression: A nationally representative cohort study. *J. Affect. Disord.* **190,** 395–406 (2016).

30. Cohen, A. S. & Bolt, D. M. A Mixture Model Analysis of Differential Item Functioning. *J. Educ. Meas.* **42,** 133–148 (2005).

31. Maij-de Meij, A. M., Kelderman, H. & van der Flier, H. Improvement in Detection of Differential Item Functioning Using a Mixture Item Response Theory Model. *Multivar. Behav. Res.* **45,** 975–999 (2010).

32. Tay, L., Newman, D. A. & Vermunt, J. K. Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organ. Res. Methods* **14,** 147–176 (2011).

33. Maij-de Meij, A. M., Kelderman, H. & Flier, H. van der. Latent-Trait Latent-Class Analysis of Self-Disclosure in the Work Environment. *Multivar. Behav. Res.* **40,** 435–459 (2005).

34. Maij-de Meij, A. M., M.|Kelderman, A., Flier, H. der & Henk. Fitting a Mixture Item Response Theory Model to Personality Questionnaire Data: Characterizing Latent Classes and Investigating Possibilities for Improving Prediction. *Appl. Psychol. Meas.* **32,** 611–631 (2008).

35. Egberink, I. J. L., Meijer, R. R. & Veldkamp, B. P. Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *J. Res. Personal.* **44,** 232–244 (2010).

36. Austin, E. J., Deary, I. J. & Egan, V. Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personal. Individ. Differ.* **40,** 1235–1245 (2006).

37. Mineka, S., Watson, D. & Clark, and L. A. Comorbidity of Anxiety and Unipolar Mood Disorders. *Annu. Rev. Psychol.* **49,** 377–412 (1998).

38. Kramer, T., Smith, G. & Maruish, M. in *The use of psychological testing for treatment planning and outcomes assessment, 3rd ed Instruments for adults* 293–311 (2004).

39. McKnight, P. E. & Kashdan, T. B. Purpose in life as a system that creates and sustains health and well-being: An integrative, testable theory. *Rev. Gen. Psychol.* **13,** 242–251 (2009).

40. Wanders, R. B. K. *et al.* Casting wider nets for anxiety and depression: disability-driven cross-diagnostic subtypes in a large cohort. *Psychol. Med.* **46,** 3371–3382 (2016).

41. Bijl, R. V., Ravelli, A. & Zessen, G. van. Prevalence of psychiatric disorder in the general population: results of the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Soc. Psychiatry Psychiatr. Epidemiol.* **33,** 587–595 (1998).

42. De Graaf, R. *et al. Response and non-response third wave: the Netherlands Mental Health Survey and Incidence Study (NEMESIS) Technical Report no. 11.* (Trimbos-Institute, 2000).

43. Graaf, R. de, Bijl, R. V., Smit, F., Ravelli, A. & Vollebergh, W. A. M. Psychiatric and Sociodemographic Predictors of Attrition in a Longitudinal Study The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Am. J. Epidemiol.* **152,** 1039–1047 (2000).

44. Stewart, A. L., Hays, R. D. & Ware, J. E. The MOS Short-Form General Health Survey: Reliability and Validity in a Patient Population. *Med. Care* **26,** 724–735 (1988).

45. Pearlin, L. I. & Schooler, C. The structure of coping. *J. Health Soc. Behav.* **19,** 2–21 (1978).

46. Rosenberg, M. *The measurement of Self-esteem*. (Princeton University Press, 1965).

47. Goldberg, D. P. & Blackwell, B. Psychiatric Illness in General Practice: A Detailed Study Using a New Method of Case Identification. *Br Med J* **2,** 439–443 (1970).

48. Muthén, L. K. & Muthén, B. O. *Mplus User's Guide. 5th ed.* (Muhtén & Muthén, 1998).

49. Vermunt, J. & Magidson, J. *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. (Statistical Innovations Inc., 2013).

50. Clark, S. *et al.* Models and strategies for factor mixture analysis: Two examples concerning the structure underlying psychological disorders. *Struct. Equ. Model.* **20,** 681–703 (2013).

51. Shafer, A. B. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *J. Clin. Psychol.* **62,** 123–146 (2006).

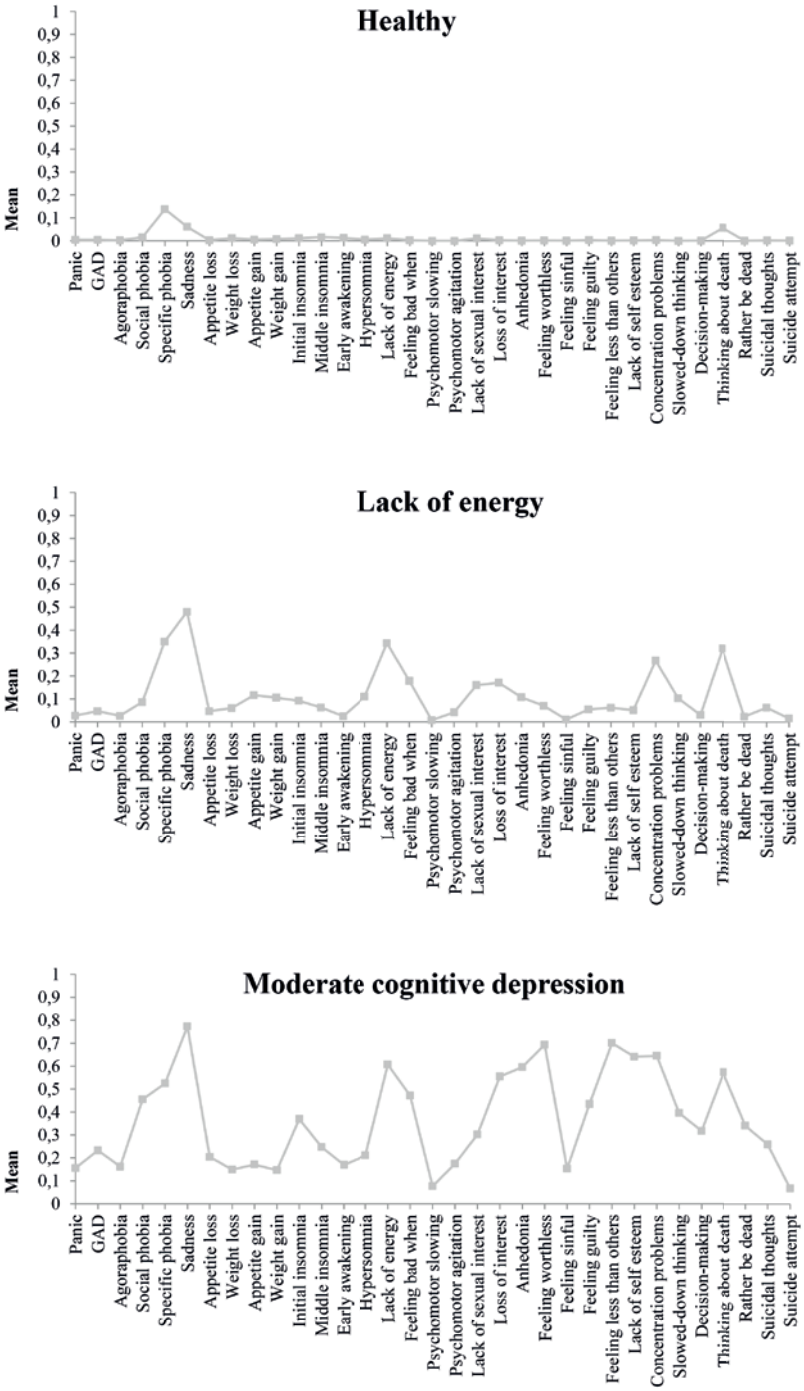
52. Wardenaar, K. J. *et al.* The structure and dimensionality of the Inventory of Depressive Symptomatology Self Report (IDS-SR) in patients with depressive disorders and healthy controls. *J. Affect. Disord.* **125,** 146–154 (2010).

53. Wardenaar, K. J. Syndromes versus symptoms : towards validation of a dimensional approach of depression and anxiety. (2012).

54. Monden, R., Wardenaar, K. J., Stegeman, A., Conradi, H. J. & Jonge, P. de. Simultaneous Decomposition of Depression Heterogeneity on the Person-, Symptom- and Time-Level: The Use of Three-Mode Principal Component Analysis. *PLOS ONE* **10,** e0132765 (2015).

55. Wardenaar, K. J., Monden, R., Conradi, H. J. & de Jonge, P. Symptom-specific course trajectories and their determinants in primary care patients with Major Depressive Disorder: Evidence for two etiologically distinct prototypes. *J. Affect. Disord.* **179,** 38–46 (2015).

SUPPLEMENT 1. Mean item scores in each of the classes of the 6-class latent class model



Sleep problems
Mean
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Panic
GAD
Agoraphobia
Social phobia
Specific phobia
Sadness
Appetite loss
Weight loss
Appetite gain
Weight gain
Initial insomnia
Middle insomnia
Early awakening
Hypersomnia
Lack of energy
Feeling bad when
Psychomotor slowing
Psychomotor agitation
Lack of sexual interest
Loss of interest
Anhedonia
Feeling worthless
Feeling sinful
Feeling guilty
Feeling less than others
Lack of self esteem
Concentration problems
Slowed-down thinking
Decision-making
Thinking about death
Rather be dead
Suicidal thoughts
Suicide attempt

Moderate somatic depression
Mean
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Panic
GAD
Agoraphobia
Social phobia
Specific phobia
Sadness
Appetite loss
Weight loss
Appetite gain
Weight gain
Initial insomnia
Middle insomnia
Early awakening
Hypersomnia
Lack of energy
Feeling bad when
Psychomotor slowing
Psychomotor agitation
Lack of sexual interest
Loss of interest
Anhedonia
Feeling worthless
Feeling sinful
Feeling guilty
Feeling less than others
Lack of self esteem
Concentration problems
Slowed-down thinking
Decision-making
Thinking about death
Rather be dead
Suicidal thoughts
Suicide attempt

Severe depression
Mean
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Panic
GAD
Agoraphobia
Social phobia
Specific phobia
Sadness
Appetite loss
Weight loss
Appetite gain
Weight gain
Initial insomnia
Middle insomnia
Early awakening
Hypersomnia
Lack of energy
Feeling bad when
Psychomotor slowing
Psychomotor agitation
Lack of sexual interest
Loss of interest
Anhedonia
Feeling worthless
Feeling sinful
Feeling guilty
Feeling less than others
Lack of self esteem
Concentration problems
Slowed-down thinking
Decision-making
Thinking about death
Rather be dead
Suicidal thoughts
Suicide attempt

SUPPLEMENT 2. Posterior class probabilities for different scoring levels on the covariates, which were included in the 3-class MM-IRT-C model.

SF-36 scale	Score levels	Posterior class-probabilities		
		Healthy	Mood/Anhedonia	Sleep/Energy
Social functioning	0-77.5%	0.43	0.32	0.25
	80-90%	0.72	0.14	0.14
	100%	0.86	0.05	0.09
Physical functioning	0-85%	0.58	0.19	0.22
	90-95%	0.70	0.15	0.16
	100%	0.81	0.09	0.09
Physical role functioning	0 (<100%)	0.54	0.24	0.22
	1 (100%)	0.79	0.10	0.11
Emotional role functioning	0 (<100%)	0.30	0.45	0.25
	1 (100%)	0.80	0.08	0.12

CHAPTER 9

Person-fit Feedback on Inconsistent Symptom Reports in Clinical Depression Care

Rob B. K. Wanders, Rob R. Meijer,
Henricus G. Ruhé, Sjoerd Sytema,
Klaas J. Wardenaar, Peter de Jonge

Under review

ABSTRACT

Objective. To evaluate the validity of person-fit statistics to identify inconsistent symptom reporting and to assess the clinical usefulness of providing clinicians with person-fit score feedback during depression assessment.

Method. Inconsistent response behavior on the Inventory of Depressive Symptomatology Self-Report (IDS-SR) was investigated quantitatively with person-fit statistics for both intake and follow-up measurements in the Groningen University Center of Psychiatry (n=2036). Subsequently, to investigate the causes and clinical usefulness of on-the-fly person-fit alerts, qualitative follow-up assessments were conducted with three psychiatrists about twenty of their patients that were randomly selected.

Results. Inconsistent symptom profiles at intake (12.3%) were predominantly characterized by reporting of severe symptoms (e.g., psychomotor slowing) without mild symptoms (e.g., irritability). Person-fit scores at intake and follow-up were positively correlated (r=0.45). Qualitative interviews with psychiatrists resulted in an explanation for the inconsistent response behavior (e.g., complex comorbidity, somatic complaints) for 19 of 20 patients. Psychiatrists indicated that if provided directly after the assessment, a person-fit alert would have led to new insights in 60%, and reason for discussion with the patient in 75% of the cases.

Conclusion. Providing clinicians with automated feedback when inconsistent symptom profiles occur is informative and can be used to support clinical decision-making.

INTRODUCTION

Psychiatrists and psychologists make extensive use of questionnaires in clinical decision-making. The use of item response theory (IRT[1]) models to construct and to evaluate the psychometric quality of such clinical questionnaires is becoming the standard procedure. The application of these models enables clinical researchers to construct computer adaptive tests, to detect differential item and test functioning, and to assess unexpected test behavior in individuals. Studies focused on this latter topic are grouped under the banner of person-fit research[2].

The aim of person-fit research is to detect item-score patterns that are inconsistent compared to the other patterns in the sample, or that are inconsistent given the test model that is assumed to describe the data[3,4]. Inconsistent response patterns may result in invalid test scores, or at least test scores that are difficult to interpret. For example, unmotivated respondents may more or less randomly fill out a questionnaire, or severely depressed patients may be inconsistent with respect to their true state because they would like to hide certain symptoms (e.g. to prevent unwanted hospitalization). Detecting such inconsistent response patterns may have diagnostic value. Furthermore, detecting invalid test scores may have value in routine outcome measurement where test scores are compared across different test occasions and invalid test scores may provide wrong impressions about the patient's stability or change across test administrations.

Although many statistics are available to assess person fit, there are almost no studies that take the next logical step of examining the reason of this unexpected test behavior (see Meijer et al.[5] for an exception in the educational field) and how this behavior can invalidate test interpretation by a clinician. Embretson and Reise[6] noted that *"there is scant evidence that person-fit scores mean anything psychologically about an individual, or are even useful for invalidating scale scores. That is psychometric researchers seem especially adept at creating "new" person fit indices, or "cleaning up" their sampling distributions, but fumble when it comes to studying their validity"*. Indeed, there are many theoretical studies concerning person-fit measurement but we know of no study where person-fit scores are incorporated into the assessment process to evaluate their practical merits.

AIMS OF THE STUDY

To our knowledge, this is the first study in the clinical field to investigate if there is some psychological reality behind inconsistent or misfitting item-score patterns. The study was conducted in a psychiatric setting, where a computer based automatic test-administration system was used to assess depression severity in patients that presented to a specialized

mental health care institute. Using data from this system, person-fit statistics were calculated for 2036 patients and those with inconsistent response patterns were identified. First, the content validity of these flagged item response patterns was evaluated. Next, to evaluate the possible usefulness of an on-the-fly feedback system for caregivers, we asked three psychiatrists if, in retrospect, an ”inconsistency alert” would have been helpful for them in the interpretation of their patients’ depression severity scores.

MATERIAL AND METHODS

PARTICIPANTS AND PROCEDURE

Data came from 2036 patients who completed at least one Inventory of Depressive Symptomatology Self-Report (IDS-SR) questionnaire (see below) in 2014 at the University Center of Psychiatry in Groningen. The IDS-SR was used both as an instrument to assess the baseline severity of depressive symptoms at intake, as well as to monitor the change in severity over time. Using an online routine outcome monitoring system called RoQua (www.roqua.nl), patients were invited to complete questionnaires before, during, and after treatment. Using a personal login code, patients could complete the questionnaires at home or at the University Center of Psychiatry, where support was available when necessary. Patients were informed that their anonymized data could be used for research purposes prior to data collection. Because anonymized, existing data was used and our study did not involve interventions, the medical ethics committee of the University Medical Center Groningen waived formal judgement of the study.

After the completion of a questionnaire by the patient, the clinicians obtained a feedback report and could further inspect responses to individual items. For the IDS-SR the feedback report consisted of the total score and a corresponding severity indicator based on cutoff scores. Person-fit statistics were implemented within this feedback report in the RoQua system (Supplement 1), and data was retrospectively extracted (see below). Whenever a patient completed an IDS-SR assessment, besides the regular feedback an extra alert was visible informing the clinician about possible inconsistencies within the reported IDS-SR symptom pattern based on the results of the person-fit statistic. An extensive explanation of the person-fit alert was given, together with a warning that the alert should be used for research purposes only.

Data for the current study came from the assessment at intake, as well as from the repeated measures during and after treatment. As a result, the number of measurements varied across patients. Of the 2036 patients, 754 (37.0%) patients were assessed repeatedly. Of them, 273 (13.4%) had two measurements and 481 (23.6%) had more than two measurements, providing a total of 6091 IDS-SR response vectors. The patient data

came from 104 clinicians, and included data on age, gender and clinical diagnosis in addition to the IDS-SR measurement. Some clinicians used the IDS-SR more frequently than others, with 6 psychiatrists having more than 200 IDS-SR entries and more than 88 patients each, and 31 psychiatrists having less than 5 patients with any IDS-SR entries in the system. Of the patients in the sample, 1021 (50.1%) completed the IDS-SR at the outpatient clinic for general psychiatry, and 1015 (49.9%) completed the IDS-SR as patients in specialized psychiatric care programs.

As a first pilot to evaluate the validity and potential usefulness of implemented person-fit alerts, we conducted a qualitative follow-up study. Three psychiatrists were asked in retrospect to give detailed feedback about a total of twenty of their patients that were randomly selected from all their patients that were flagged as inconsistent responders. Feedback was obtained through a questionnaire (Supplement 2) that contained closed and open questions, both about the nature of the inconsistency as well as the possible usefulness of such an alert for clinical practice.

INSTRUMENT

The Inventory of Depressive Symptomatology Self-Report (IDS-SR[7]) is a self-report questionnaire consisting of 30 items rated on a 4-point (0-3) Likert scale to assess the severity of depressive symptoms. As a patient could either endorse 'appetite increase' or 'appetite decrease' and either 'weight increase' or 'weight decrease', these items were combined respectively into compound 'appetite change' and 'weight change' items. The IDS-SR assesses all DSM-IV criterion symptoms for Major Depressive Disorder (MDD) and the most commonly associated non-criterion symptoms (e.g., anxiety, irritability).

STATISTICAL ANALYSIS

Person fit

Person-fit statistics enable the identification of patients for whom the observed symptom pattern is different than would be expected based upon the model used to describe the data[2]. For a patient with some level of depression severity, the probability of reporting depressive symptoms is expected to decrease as symptoms become more severe. That is, milder symptoms are more likely to be reported than more severe symptoms. The more a symptom profile deviates from this pattern of decreasing endorsement with increasing severity, the poorer person-fit will be. In many such cases misfit is caused by more severe symptoms being reported (e.g. suicidal ideation) without the reporting of milder symptoms (e.g. sad mood).

In the current study, person-fit analyses were performed using the likelihood-based standardized l_z statistic[8] with the graded response model (GRM[9]) as IRT model describing the data. This model was chosen because of the ordinal nature of the IDS-SR response data. Two parameters are estimated under the GRM model for each item: the discrimination parameter (α) reflects how strong a symptom (item) is related to underlying depression severity, and the threshold (β) is reflective of symptom severity. The IRT model was calibrated on a sample of depressed and anxiety patients, who completed the IDS-SR (n=2,329 [10]) in the Netherlands Study of Depression and Anxiety (NESDA[11]).

It was chosen to calibrate the IRT model on a well-defined external sample as the current sample of psychiatric patients consisted of a heterogeneous group with very diverse psychopathology, and fitting an IRT model in such a heterogeneous sample might result in interpretation problems. The NESDA sample is well defined in terms of psychopathology through the use of a standardized structured CIDI interview, and the use of person-fit statistics was previously investigated in this sample[10]. Since the purpose of the IDS-SR is to measure depression severity in patients with depression, using the IRT model from the well-defined calibration sample in the person-fit analyses allowed for identification of patients, for whom this model did not hold. In these analyses, the l_z statistic then represents how likely it is to observe a patient's reported symptom pattern given this model of depressive symptoms. Patients with higher values of l_z have symptom patterns consistent with the IRT model, whereas lower values of l_z indicate poor person-fit and represent patients with symptom profiles that are inconsistent with the model, and therefore not a good reflection of depression severity.

Analyses

First, person-fit analyses were performed on the intake IDS-SR assessments of all patients (n=2036). Patients were divided into two groups based on their person-fit score compared to a 5% significance level cutoff ($l_z < -1.39$) and to a 1% significance level cutoff ($l_z < -2.21$), obtained from the reference person-fit study in the calibration sample[10]. Patients with person-fit scores below the cutoff score were allocated to an 'inconsistent group' and were further investigated in terms of symptom patterns and associated external variables.

Second, stability of person-fit scores across repeated measurements was investigated in 754 patients, who had follow-up assessments. The correlation between person-fit scores on the first and second measurement, and the proportion of measurements that were flagged as inconsistent at each measurement were investigated. In addition, the response patterns of six patients with more than 18 measurements were randomly selected and investigated in more detail. Here, two patients were selected with no measurements flagged as inconsistent, two were selected with less than 25% flagged as inconsistent, and two were selected with more than 90% of their measurements flagged as inconsistent.

Third, results of qualitative follow-up assessments on twenty randomly selected patients with poor person-fit scores of three psychiatrists were investigated. Both explanations of psychiatrists on the potential causes of poor person-fit, as well as the potential clinical usefulness of a person-fit alert for psychiatrists at the time of actual measurement were retrospectively analyzed.

RESULTS

DESCRIPTIVES

The sample had a mean age of 43.6 years (s.d.=14.5) and included 1061 women (52.1%; Table 1). Patients showed a wide variety of diagnoses, with most patients diagnosed with a primary clinical diagnosis of a mood disorder (25.1%), or anxiety disorder (15.6%). A secondary clinical diagnosis was observed for 33.6% of the patients of which 517 (25.4%) had a secondary Axis I diagnosis (e.g. a comorbid anxiety disorder) and 167 (8.2%) had a secondary Axis II diagnosis (comorbid personality disorder).

TABLE 1. Descriptives of UCP patient sample (N=2036)

Characteristic	Mean or Frequency	SD or %
Male gender	975	(47.9%)
Age	43.6	(14.5)
IDS-SR score	26.9	(15.0)
IDS-SR measurements	3.0	(5.0)
Primary clinical diagnosis		
Anxiety disorder	318	(15.6%)
Bipolar disorder	195	(9.6%)
Childhood and developmental disorder	167	(8.2%)
Mood disorder	512	(25.1%)
MDD - First episode	162	(8.0%)
MDD - Recurrent	309	(15.2%)
Other	41	(2.0%)
Personality disorder	120	(5.9%)
Schizofrenia or psychotic disorder	57	(2.8%)
Somatoform disorder	77	(3.8%)
Other clinical disorder	119	(5.8%)
Secondary clinical diagnosis		
Axis I disorder	517	(25.4%)
Axis II disorder	167	(8.2%)

SD, Standard Deviation; IDS-SR, Inventory of Depressive Symptomatology-Self Report; MDD, Major Depressive Disorder.

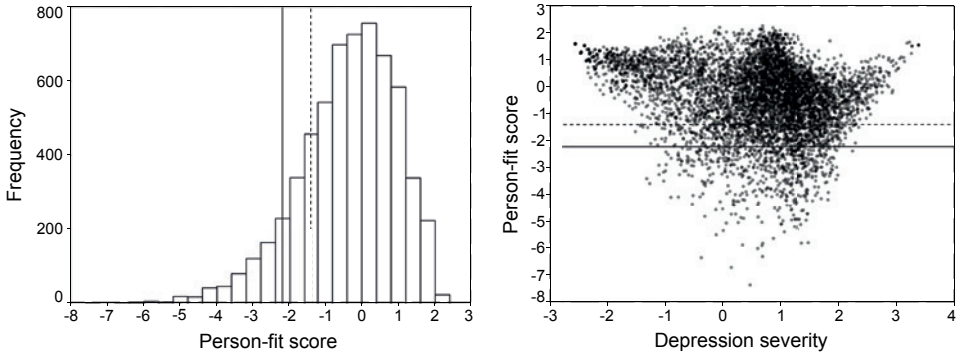


FIGURE 1. Person-fit score distribution (left) and across different levels of depression severity (right) with the dotted line representing the 5% cutoff-score (l_z <-1.39) and the solid line the 1% cutoff-score (l_z <-2.21).

PERSON-FIT AT INTAKE

Person-fit analyses were first performed on the IDS-SR assessments of all patients at intake. The distribution of person-fit scores (mean l_z = -0.41, sd l_z = 1.35) was skewed to the left (Figure 1), and showed a distinct V-shape across different levels of depression, with higher person-fit scores on the extreme ends of the depression severity spectrum. However, there was no relation between person-fit and depression severity (ρ=0.03; 95% CI [-0.02,0.08]) in those with poor person-fit scores (l_z < -1.39).

Of all patients at intake (n=2036), 543 (26.6%) had person-fit scores below the 5% significance level and 260 (12.8%) had person-fit scores below the 1% significance level. In patients with a primary mood diagnosis (n=512), person-fit scores were below the 5% significance level for 122 (23.8%), and below the 1% significance level for 63 (12.1%) patients. For further analyses, the more conservative 1% significance level was taken as cut-off for poor person-fit, indicative of inconsistent response behavior.

Person-fit scores for the first measurement wave were further investigated in terms of symptom profiles and associated external variables. Mean item scores of the response patterns flagged as inconsistent were substantially different from the mean item score patterns in the typical responders (Supplement 3). Inconsistent response patterns were characterized by lower scores on 'anxious', 'somatic complaint', 'sympathetic arousal', and 'sensitivity' and higher levels of 'reactivity of mood', 'involvement', 'enjoyment', and 'psychomotor slowing' (all mean differences were significant at p<0.001). Multivariate logistic regression showed that inconsistent patterns were more often observed in older patients (Cohen's d=0.17; z=2.74; p<0.01), male patients (OR=1.6; z=3.41; p<0.01), and in patients without a primary clinical diagnosis of an anxiety disorder (OR=0.6; z=-2.34; p<0.05). Presence of other clinical diagnoses was not a significant predictors of inconsistent response patterns.

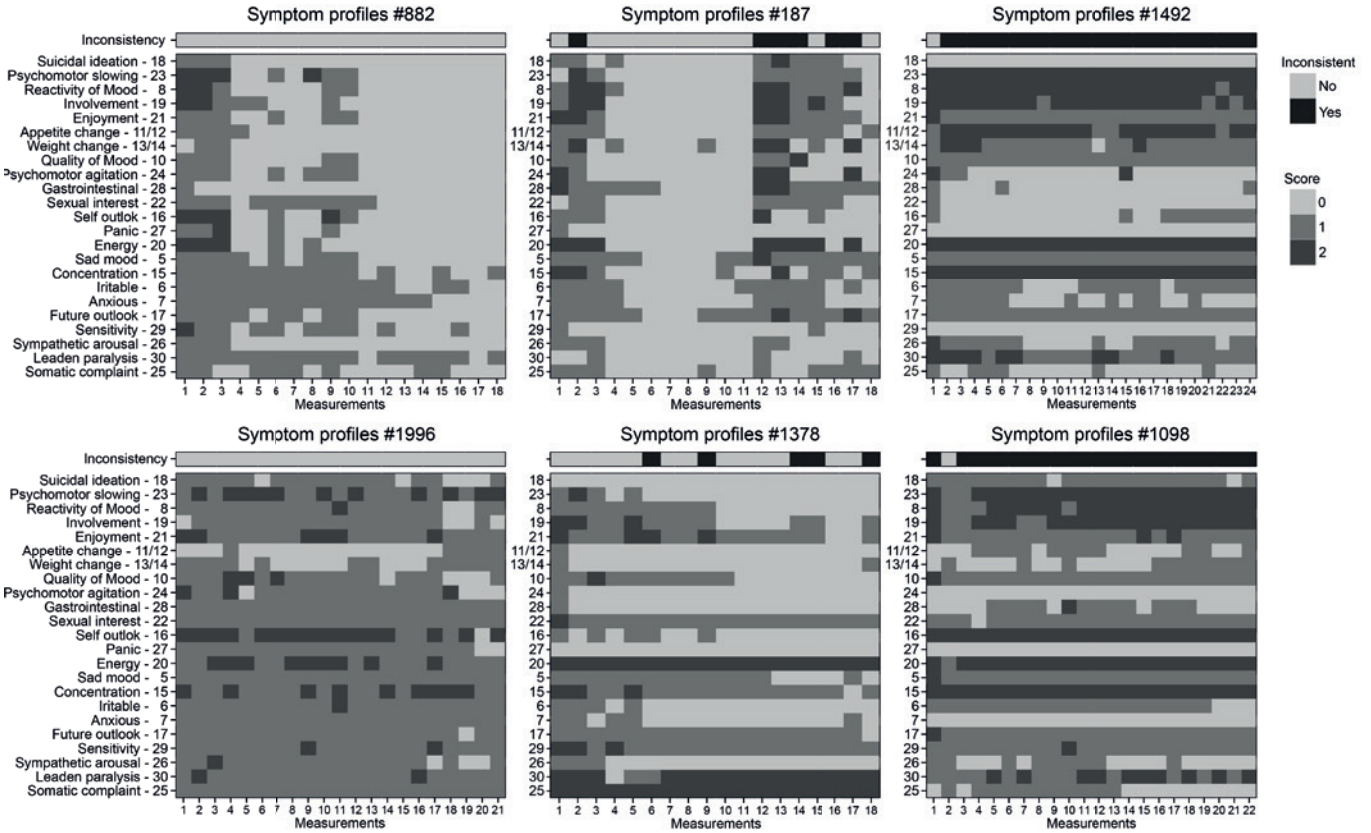


FIGURE 2. Symptom profiles of six patients with repeated measurements with depressive symptoms ordered from mild (bottom) to severe (top) based on severity thresholds obtained from the IRT model. Each block represents a score (0,1,2) reported on the symptom at the time of measurement. For the first two patients on the left (#1996 and #892) no measurements are flagged as inconsistent, for the next two patients in the middle (#1378 and #187) <25% are flagged as inconsistent, and the last two patients on the right (#1492 and #1098) have >90% measurements flagged as inconsistent.

PERSON-FIT ON REPEATED MEASUREMENTS

Person-fit scores at the first measurement were positively correlated with person-fit scores on the second measurement (r=0.45). Interestingly, several patients had stable inconsistent profiles over multiple measurements (Supplement 4). Anecdotally, one patient (#1492) with a primary diagnosis of MDD (first episode) had 24 measurements, of which 21 were flagged as inconsistent at the 1% significance level (person-fit range: -1.1 to -5.1; mean l_z =-3.12).

To gain more insight into the possible consistency of inconsistent response behavior across multiple measurements, the repeated measurements of six randomly selected patients are plotted in Figure 2. The two patients (#1492 and #1098) with more than 90% measurements flagged as inconsistent showed similar inconsistencies across all measurements, suggesting a systematic cause underlying the reporting of a symptom pattern not reflective of depression. Patterns of both patients (Figure 2) showed high scores on severe symptoms (e.g. 'Psychomotor slowing' and 'Reactivity of mood') while many mild symptoms were not reported (e.g. 'Involvement' or 'Sensitivity'). Both examples also suggest that poor person-fit is not simply caused by a single reporting of a severe symptom without reporting of milder symptoms, but instead poor person-fit scores are observed when many of these deviations are present.

Alternatively, patient #1378 showed a pattern where the first measurements were not flagged as inconsistent but later measurements were (especially the last measurement with a person-fit score of -2.3). At the first measurements, many symptoms are reported, with an IDS sum score of 49 at intake. In later measurements, depressive severity improved, with a total score of 24 on measurement 10 (5 months later). At the remaining 8 measurements the sum score remained around 24, with the patient reporting mainly symptoms of 'Enjoyment', 'Sexual Interest', 'Energy', 'Somatic complaint', and 'Leaden paralysis'. This suggests that although the depression improved overall, residual symptoms remained that led to a potentially overestimated depression severity at later measurements.

QUALITATIVE FOLLOW-UP ASSESSMENTS ON PERSON-FIT

Psychiatrists explanations on potential causes of low person-fit

Summaries of qualitative assessments where three psychiatrists were asked on the potential causes of inconsistent symptom profiles for twenty of their patients with poor person-fit are given in Table 2. The randomly selected patients had an average age of 46.2 (range: 21-80), showed mild to severe depression severity with an average IDS sum score of 36 (range: 17-52) and showed poor person-fit scores averaging -3.2 (range: -2.2 to -4.7).

As some patients may have filled out an IDS while not under treatment of the psychiatrist, the psychiatrists were first asked how well they were acquainted with the patient.

Psychiatrists reported that they were well acquainted with the patient in 17 of the 20 cases, and reasonably well acquainted with the patient in the remaining three cases. Psychiatrists reported that for 14 of the 20 patients the inconsistent response behavior conformed to the clinical impression they had of these patients. In most cases the explanation given for the inconsistent behavior was that the IDS profiles contained symptoms experienced for other reasons than MDD. For example, psychiatrists mentioned complex comorbidity (e.g. #379), somatic complaints (e.g. #1378), or the presence of isolated symptoms (e.g. #1975) as possible explanations. For six patients, the inconsistent response behavior was in retrospect not in agreement with the psychiatrists' clinical impressions. Here, an alternative explanation could be offered in five cases. These explanations pointed at high levels of psychiatric distress with severe problems (e.g. #2898), and motivational or concentration problems (e.g. #1723). Overall, these results showed that poor person-fit could be linked to a diverse range of possible underlying causes. To illustrate this, three cases are discussed in more detail below.

Patient #926 had a depression with somatic comorbidity (as reported by the psychiatrist) resulting in a depressive symptom profile with high scores on somatic symptoms, reporting for example gastrointestinal problems and low energy. These symptoms were not reflective of depression severity and lead to deviations from the typical response pattern with an inflated total score as a result. This example shows that information obtained from person-fit scores may have added value above the total score and that a psychiatrist should be careful when interpreting the total score.

Interestingly, for patient #1531 the explanation of the psychiatrist was that the patient had a wish to be discharged and presumably might have pretended to be better than he actually was. This behavior might have resulted in an inconsistent pattern with poor person-fit, since the patient may have reported improvement of some obvious depressive symptoms and not on others. Inspection of the two IDS measurements prior to the inconsistent one showed extremely high IDS scores of 57 and 64 (indicative of severe depression in the past month, both with good person-fit l_z >0.7), strengthening the interpretation of the psychiatrist of a possible under-estimation of depression severity at the inconsistent measurement (IDS score of 17, see Table 2).

Patient #187 reported few symptoms and only had a moderate IDS score of 26, but did report severe symptoms like psychomotor agitation and psychomotor slowing, which resulted in a poor person-fit score. The patient showed abnormalities during neurological examination, with problems in information processing and slowness of thought (bradyphrenia). For this patient the IDS score may therefore not be reflective of depression symptomatology (i.e. severity) at all but rather of neurological defects, showing the potential extra information that an alert of poor person-fit could have for a clinician.

TABLE 2. Explanation of psychiatrists for the inconsistency of reported symptom patterns of twenty of their patients.

ID	Age	l_z	IDS	Familiar with patient	Conform clinical impression[1]	Severity estimation[2]	Explanation psychiatrist on potential cause
187	26	-4.3	26	Reasonable	Yes	-	Problems in information processing, bradyphrenia, abnormalities in neurological examination and referred to neurologist.
379	43	-3.2	41	Very good	Yes	Under	Depression with complex comorbidity causing very high level of suffering at time of measurement.
575	39	-3.8	39	Reasonable	-	Good	Presence of isolated symptoms
766	48	-3.2	48	Very good	Yes	Good	Severe depression with complex comorbidity including anxiety and dissociation.
837	21	-3.5	21	Good	No	Good	No explanation.
926	43	-4.2	43	Good	Yes	Over	Somatic comorbidity, many physical complaints especially pain.
1339	41	-3.5	41	Very good	Yes	Under	Motivation/concentration problems in an episode of severe decompensation.
1378	51	-2.3	24	Good	Yes	Over	Many somatic complaints leading to a higher score than expected based on patient's mood.
1487	51	-2.8	52	Good	Yes	Over	High psychiatric distress with comorbid anxiety and catastrophic interpretation of pain symptoms, causing patient to be desperate about the future and suicidal.
1531	71	-2.2	17	Good	No	Under	Patient wanted to be discharged and might have pretended to be better, although patient showed some clinical improvement in the week before.
1543	49	-2.4	28	Very good	Yes	Under	Patient inexplicably improved during a wash-out phase (no medication), which was surprising as the patient still seemed mentally unstable and quite ill.
1704	52	-3.2	48	Good	Yes	Under	Patient has bipolar depression with more psychomotor retardation and energy loss. Possibly, [the patient] was more depressed than reflected by the questionnaire as [the patient] showed limited illness awareness.
1723	28	-2.4	28	Good	No	Under	Motivation/concentration problems.
1975	49	-3.2	50	Very good	Yes	Good	Presence of isolated symptoms and high psychiatric distress.
2541	28	-3.8	28	Good	Yes	Over	Motivation/concentration problems.
2775	48	-3.9	32	Reasonable	Yes	Good	Comorbid autism spectrum disorder (PDD-NOS).
2816	80	-4.7	35	Very good	Yes	Under	Patient was treated with eskatamine in a terminal phase and later deceased through euthanasia as a result of total despair. Patient showed a tendency to trivialize his depression.
2898	57	-2.5	38	Very good	No	Under	High psychiatric distress with comorbid anxiety disorder.
3049	52	-2.5	47	Very good	Yes	Over	Exaggerates or feigns symptoms.
3058	47	-3.3	39	Very good	No	Good	High psychiatric distress

[1]Psychiatrists were asked 'Does the inconsistency alert correspond with your own clinical impression of the patient?'
[2]Psychiatrists were asked 'Was the total score an under-, over-, or good estimation of the severity of depression?'

TABLE 3. Clinical usefulness of the person-fit alert according to psychiatrists regarding twenty of their patients with inconsistent symptom patterns.

ID	Age	l_z	IDS	Useful alert[1]	New insights[2]	Inspect item scores	Discuss with patient	Explanation psychiatrist on clinical usefulness
187	26	-4.3	26	No	No	No	No	No explanation given [patient was referred to neurologist].
379	43	-3.2	41	Yes	Yes	Yes	Yes	Clinically useful, if it becomes clear how this was manifested in the response behavior.
575	39	-3.8	39	No	Yes	Yes	Yes	Patient came only for specific chronotherapy and an alert of inconsistency would not have led to changes in [the used treatment] policy. Would have led to a further discussion with patient.
766	48	-3.2	48	Yes	Yes	Yes	Yes	Could be helpful, if it would provide more clarity about the nature of the inconsistency
837	21	-3.5	21	Yes	Yes	Yes	Yes	Possibly insightful, [psychiatrist says] to be curious on which domains the inconsistency occurred.
926	43	-4.2	43	No	No	Yes	No	Inconsistency was already expected.
1339	41	-3.5	41	Yes	No	Yes	Yes	Would have led to further discussion with patient. [Psychiatrist says] the inconsistency fits within the clinical picture of severe problems.
1378	51	-2.3	24	No	No	No	No	The inconsistent response behavior was expected, as to us the patient had shown good clinical improvement in mood.
1487	51	-2.8	52	Yes	Yes	Yes	Yes	Depending on where the inconsistency is found; if the patient is very suicidal and anxious but has lower depression severity, [the report] would fit with [the psychiatrist's] impression and could help to interpret the high IDS score, which [the psychiatrist] finds clinically incorrect.
1531	71	-2.2	17	Yes	Yes	Yes	Yes	Depending on why the measurement was inconsistent it could lead to further discussion with the patient especially given his strong wish to be discharged.
1543	49	-2.4	28	Yes	Yes	Yes	Yes	Could have provided more insight. Patient later deteriorated and this could perhaps been detected as a result of the person-fit alert.
1704	52	-3.2	48	No	No	No	No	Patient was clinically clearly ill.
1723	28	-2.4	28	No	No	No	No	It was clear that this measurement was aberrant since other measures were considerable higher, and this corresponded better with our observations during treatment.
1975	49	-3.2	50	Yes	Yes	Yes	Yes	[Psychiatrist says] it would make him alert and is reason for further discussion. Possibly, [psychiatrist] missed something.
2541	28	-3.8	28	Yes	Yes	Yes	Yes	Possibly insightful.
2775	48	-3.9	32	No	No	No	No	Fits within the clinical picture of comorbidity.
2816	80	-4.7	35	Yes	Yes	No	Yes	Help to increase understanding, and could have been a reason to discuss despair and the downplaying of his depression
2898	57	-2.5	38	Yes	Yes	Yes	Yes	Could lead to further diagnostic examination.
3049	52	-2.5	47	Yes	Yes	Yes	Yes	Would strengthen the current clinical picture.
3058	47	-3.3	39	Yes	No	Yes	Yes	Would confirm that the severity indeed is high.

[1]Psychiatrists were asked: 'Would the alert of possible inconsistency be useful in this case for you as a psychiatrist?'
[2]Psychiatrists were asked: 'Could the alert of possible inconsistency have led to new insights?'

Clinical usefulness of person-fit alerts

Psychiatrists were asked about the potential clinical usefulness of a person-fit alert if offered at the time of the actual IDS measurement (Table 3). For 13 of the 20 twenty investigated patients in the follow-up interview, psychiatrists indicated that the alert would have been of direct clinical use. For the remaining seven patients, the alert would not have been of direct clinical use according to the psychiatrists. They reported that for five of these cases the inconsistency was already expected at the time and fitted within the clinical picture. One of these patients (#187; described above) was referred to a neurologist, and one patient (#575) was reported to have come for planned specialized treatment, for which the IDS measurement would not have led to changes in treatment policy.

Psychiatrists indicated that for 12 of the 20 patients the person-fit alert would have led to new insights. Here, it was reported that the alert could have increased understanding (#2816), alerted the psychiatrist to things they could potentially have missed (#1975) and could have led to further diagnostic examination (#2898). For the other eight patients the inconsistency was already expected and conformed to the clinical picture. Here, the psychiatrists answered that they did not expect the alert to have led to any new insights or actions. Still, the psychiatrists pointed out that the alert would have been a useful confirmation of the clinical impression that they had of the patient, and could have helped to interpret their IDS scores (e.g. patient #1487).

With regard to potential actions taken after a person-fit alert, psychiatrists reported that they would have inspected the individual item scores in 14 out of 20 patients. In addition, they reported that they would have wanted to know more about the nature of the inconsistency (e.g. #837, #1487). For 15 of the 20 patients, the alert would have been a reason to discuss possible inconsistencies with the patient or a useful starting point for a discussion with a patient on specific diagnostic issues. For example, the alert could have been a suitable starting point for a discussion with patient #2816 about his/her possible downplaying of depression severity.

DISCUSSION

This study aimed to investigate the clinical meaning and usefulness of inconsistent symptom profiles on IDS-SR depression severity assessments in a naturalistic clinical setting. Depressive symptoms reported by all patients who completed the IDS-SR in 2014 in a specialized care setting at the University Center of Psychiatry (Groningen) were investigated by means of a data-driven approach based on person-fit statistics. Inconsistent profiles of depressive symptoms as identified by poor person-fit scores were analyzed on all intake measurements and on repeated measurements to assess temporal stability. These results were followed by a qualitative study among three psychiatrists on

twenty of their randomly selected patients with inconsistent profiles, to get more insight into potential causes of inconsistent responding and to evaluate the potential clinical use of person-fit statistics for psychiatrists.

Poor person-fit was frequently observed, with 12.8% of patients identified with inconsistent profiles at a conservative 1% significance level for person-fit, and even 26.6% at the 5% level. This is higher than previous studies reporting rates of 6.8%-14% in clinical samples[10,12–14], but in line with the expectation that patients in specialized care present with diverse and complex psychopathology[15] and experience depressive symptoms for other reasons than major depressive disorder alone[16]. Furthermore, in this setting the IDS-SR was used to screen for the presence of MDD: only 25.1% of patients actually had a primary clinical diagnosis of MDD at the time of assessment, adding to the heterogeneity of the sample. As a result, it is not surprising that a relatively large part of the patients reported symptom patterns that deviate from the typical structure of depressive symptoms, making person-fit a potentially valuable source of information when trying to interpret scale scores in this clinical setting.

In contrast to some other applications of person-fit research (e.g., in educational assessment and selection), in clinical research the causes underlying poor person-fit are often of a systematic nature. Only in a minority of patients the reasons for misfit can be attributed to low concentration, unmotivated behavior or careless responding, which play an important role in other assessment contexts. In the majority of patients, there seem to be realistic clinical causes underlying their inconsistent response profiles. Here, symptoms were likely to be experienced for other reasons than just depression, causing symptom patterns that did not adhere to the expected depression profile. These other reasons may have included the presence of comorbid pathology like anxiety disorders or somatic complaints, or the presence of a completely different primary disorder, such as a severe neurological disorder. Interestingly, the systematic influence of these factors on patients' response patterns was supported by the observation that inconsistent response behavior was rather stable across multiple measurements (e.g. patient #1492). To better distinguish between different causes of inconsistent response behavior, that is: clinical explanations versus factors associated with the quality of the responses, a more objective measure of the latter would ideally be needed. A promising approach would be to look at the time spent on answering each item in computer-based assessments, and analyze the patterns of response times[17]. Careless and unmotivated responders would be expected to have unusually short or long response times, whereas those who show inconsistent response behavior due to particular clinical causes would not be expected to show unusual response times.

In the current study we distinguished between several clinical applications of person-fit based on the results of the qualitative psychiatrist assessments. First, an alert of inconsistency can serve as a warning signal for the psychiatrist that the total score should be interpreted with caution. In some cases, psychiatrists indicated to be curious on what symptoms caused the inconsistencies. Inspection of item scores combined with available clinical information, following a discussion with the patient could clarify such discrepancies. Alternatively, clinicians could be provided with additional information together with the person-fit alert. The system could adaptively respond to inconsistencies[18] and suggest additional items or questionnaires to administer. Alternatively, the system could analyze inconsistencies on the symptom level by means of a follow-up residual analysis on the differences between expected and observed item scores to detect both deviation trends across symptoms and possible sources of single-symptom misfit[4].

Second, a person-fit alert could confirm the current clinical picture that the psychiatrist has of a patient, especially in cases where a typical profile is not expected a priori. For example in the case of patient #1378, where the psychiatrist saw a patient clinically improving, but this improvement was not reflected in the IDS-SR total scores. Retrospectively the measurement was identified as inconsistent, which could have served as a confirmation for the psychiatrist's impression and could have supported the clinical decisions made.

Third, an alert of a purely statistical method could serve as a starting point for further discussion or an opportunity to discuss particular issues with a patient. For example, in case of patient #2816, where the suspicion was that the patient was downplaying his depression, the person-fit alert could have served as a starting point to discuss a topic that might be difficult to talk about without directly blaming the patient of downplaying. Another instance was patient #1975, where the psychiatrist indicated that he might have missed something and a person-fit message would have been a reason for further discussion, possibly leading to new insights. In future research, the improvement of care when this type of feedback is provided should be studied in a pragmatic cluster randomized trial, providing person-fit feedback to one experimental group of clinicians and maintaining care as usual in the other experimental group. In such trials, several outcomes could be compared, including additional diagnostic work-up, patient and clinician satisfaction, and the quality of the working-alliance. Ultimately, pragmatic trials could be used to investigate the effects of providing automated person-fit feedback on treatment outcome and cost-effectiveness of care.

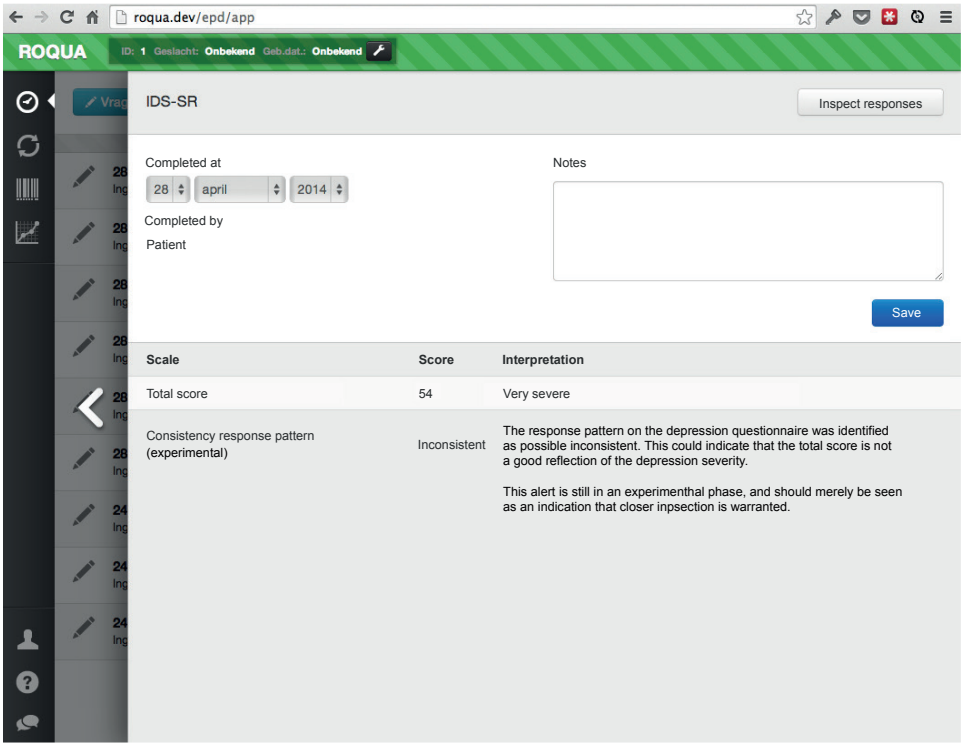
This study had several limitations. Although this is the first study in which person-fit statistics were implemented and interpreted in the context of a real clinical setting, the design was still retrospective and not prospectively based on on-the-fly feedback given in real time. Therefore, the current results should be seen as a proof-of-principle. Further prospective studies are needed, where person-fit is implemented in real time and feedback to clinicians on basis of the person-fit statistic is given on-the-fly at the actual moment of assessment. An additional limitation is that the psychiatrists knew that the person-fit scores of their patients were low before they gave their feedback (they were not blinded or provided with sham-cases of poor person-fit). In addition, we calibrated the group-based model on a well-defined but external sample (NESDA[11]). An alternative would have been to calibrate the model on only patients with a diagnosis of MDD or to obtain a more homogenous subsample by purifying the sample with the use of mixture modeling or person-fit statistics[19].

The promising results of recent person-fit studies in clinical assessment[10,12–14,20] raised important questions on the causes behind inconsistent patterns and whether implementation of on-the-fly person-fit statistics could be of clinical use. The current study affirmed that there are real clinical causes behind inconsistent symptom profiles that give poor person-fit scores a clinical interpretation. Above all, the feedback collected among psychiatrists suggested that person-fit alerts could be highly informative for clinicians when interpreting depression assessments, and of valuable support in clinical decision-making. In this clinical context, all relevant information is summarized to guide clinical decisions[21] and a person-fit message should be seen as a piece of extra information on top of the regularly used severity sum scores. With evidence converging on the usefulness of person-fit statistics, routine assessments taking place with automated systems[22], person-fit software being widely available[23,24] and non-technical tutorials being accessible[25], person-fit is ready for on-the-fly implementation in depression assessment.

REFERENCES

1. Embretson, S. & Reise, S. *Item Response Theory for Psychologists*. (Psychology Press, 2000).
2. Meijer, R. R. & Sijtsma, K. Methodology Review: Evaluating Person Fit. *Appl. Psychol. Meas.* **25,** 107–135 (2001).
3. Meijer, R. R. Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychol. Methods* **8,** 72–87 (2003).
4. Ferrando, P. in *Handbook of item response theory modeling: Applications to typical performance assessment* 128–155 (Routledge, 2015).
5. Meijer, R. R., Egberink, I. J. L., Emons, W. H. M. & Sijtsma, K. Detection and Validation of Unscalable Item Score Patterns Using Item Response Theory: An Illustration with Harter's Self-Perception Profile for Children. *J. Pers. Assess.* **90,** 227–238 (2008).
6. Embretson, S. & Reise, S. *Item Response Theory*. (Routledge/Taylor & Francis Group, (in press)).
7. Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B. & Trivedi, M. H. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol. Med.* **26,** 477–486 (1996).
8. Drasgow, F., Levine, M. V. & Williams, E. A. Appropriateness measurement with polychotomous item response models and standardized indices. *Br. J. Math. Stat. Psychol.* **38,** 67–86 (1985).
9. Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychom. Monogr. Suppl.* **34,** 100 (1969).
10. Wanders, R. B. K., Wardenaar, K. J., Penninx, B. W. J. H., Meijer, R. R. & Jonge, P. de. Data-driven atypical profiles of depressive symptoms: Identification and validation in a large cohort. *J. Affect. Disord.* **180,** 36–43 (2015).
11. Penninx, B. W. J. H. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17,** 121–140 (2008).
12. Woods, C. M., Oltmanns, T. F. & Turkheimer, E. Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychol. Assess.* **20,** 159–168 (2008).
13. Wardenaar, K. J., Wanders, R. B. K., Roest, A. M., Meijer, R. R. & De Jonge, P. What does the beck depression inventory measure in myocardial infarction patients? a psychometric approach using item response theory and person-fit. *Int. J. Methods Psychiatr. Res.* **24,** 130–142 (2015).
14. Conijn, J. M., Emons, W. H. M., De Jong, K. & Sijtsma, K. Detecting and Explaining Aberrant Responding to the Outcome Questionnaire-45. *Assessment* **22,** 513–524 (2015).
15. Groenewold, N. A. *et al.* Comparing Cognitive and Somatic Symptoms of Depression in Myocardial Infarction Patients and Depressed Patients in Primary and Mental Health Care. *PLOS ONE* **8,** e53859 (2013).
16. Wanders, R. B. K. *et al.* Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make? *J. Psychosom. Res.* **78,** 130–136 (2015).
17. Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P. & Tijmstra, J. Testing for Aberrant Behavior in Response Time Modeling. *J. Educ. Behav. Stat.* **39,** 426–451 (2014).
18. Liu, M. & Yu, P. Aberrant Learning Achievement Detection Based on Person-fit Statistics in Personalized e-Learning Systems. *Educ. Technol. Soc.* **14,** 107–120 (2011).
19. Rupp, A. A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychol. Test Assess. Model.* **55,** 3–38 (3013).
20. Conrad, K. J. *et al.* Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug Alcohol Depend.* **106,** 92–100 (2010).
21. Puschner, B. *et al.* Clinical Decision Making and Outcome in Routine Care for People with Severe Mental Illness (CEDAR): Study protocol. *BMC Psychiatry* **10,** 90 (2010).
22. Lambert, M. J. & Shimokawa, K. Collecting client feedback. *Psychotherapy* **48,** 72–79 (2011).
23. Tendeiro, J., Meijer, R. & Niessen, A. PerFit: An R package for person-fit analysis in IRT. *J. Stat. Softw.* (2015).
24. Ferrando, P. J., Lorenzo, U., Ferrando, P. J. & Lorenzo, U. WPerfit: A Program for Computing Parametric Person-Fit Statistics and Plotting Person Response Curves. *Educ. Psychol. Meas.* **60,** 479–87 (2000).
25. Meijer, R. R., Niessen, A. S. M. & Tendeiro, J. N. A Practical Guide to Check the Consistency of Item Response Patterns in Clinical Research Through Person-Fit Statistics: Examples and a Computer Program. *Assessment* **23,** 52–62 (2016).

SUPPLEMENT 1. Screenshot of an implemented person-fit alert in the online routine outcome monitoring system (*www.roqua.nl*) for the IDS-SR (translated to English). Note that the person-fit alert was not studied real-time, but data was extracted retrospectively and the three psychiatrists that were qualitatively followed-up were also asked in retrospect on causes and potential usefulness on their identified patients.



SUPPLEMENT 2. Questionnaire used to assess the potential causes and clinical uses of a person-fit alert for the patient identified with inconsistent depressive symptom pattern (translated to English). Psychiatrists were asked to give detailed explanations to each answer (see Tables 2 and 3 in the manuscript).

Person-fit questionnaire

For the current patient, the person-fit statistic identified the response pattern on the depression questionnaire as possible inconsistent. This could indicate that the total score is not a good reflection of the depression severity. We would like to find out what the causes of the inconsistent response pattern could be, by means of a short questionnaire.

Could you explain your answers as clear as possible? For this purpose, could you look up the corresponding patient records, and the completed IDS-SR in Roqua?

Patient: #ID#
Completion date of potential inconsistent IDS-SR: #date#

How familiar are you with the patient?

☐ Very good ☐ Good ☐ Reasonable ☐ Poor ☐ Very poor

Does the inconsistency alert correspond with your own clinical impression of the patient?

☐ Yes ☐ No

Was the total score an under-, over-, or good estimation of the severity of depression?

☐ Underestimation ☐ Overestimation ☐ Good estimation

What could be the potential cause of the inconsistent symptom pattern?

☐ Motivation/concentration problems ☐ Presence of isolated symptoms
☐ Exaggerates or feigns symptoms ☐ High psychiatric distress ☐ Comorbidity

Please give an extensive explanation

The following questions regard the clinical usefulness of an alert of possible inconsistency in a real clinical setting, at the time when the patient completed the IDS measurement.

Would the alert of possible inconsistency be useful in this case for you as a psychiatrist?

☐ Yes ☐ No

Please give an extensive explanation:

Could the alert of possible inconsistency have led to new insights?

☐ Yes ☐ No

Please give an extensive explanation:

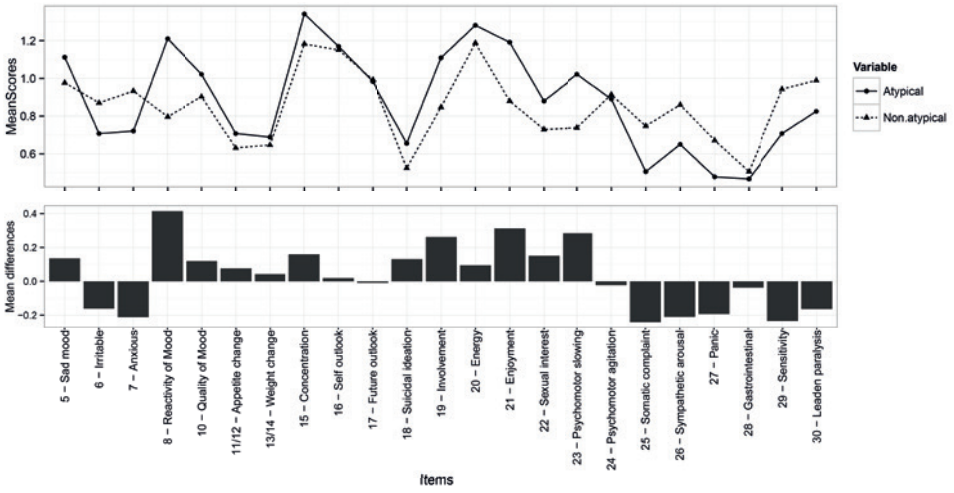
If you got the inconsistency alert at the time of measurement for this patient, what actions would you have taken?

☐ Nothing ☐ Inspect item scores ☐ Discuss with patient ☐ Inspect patient record

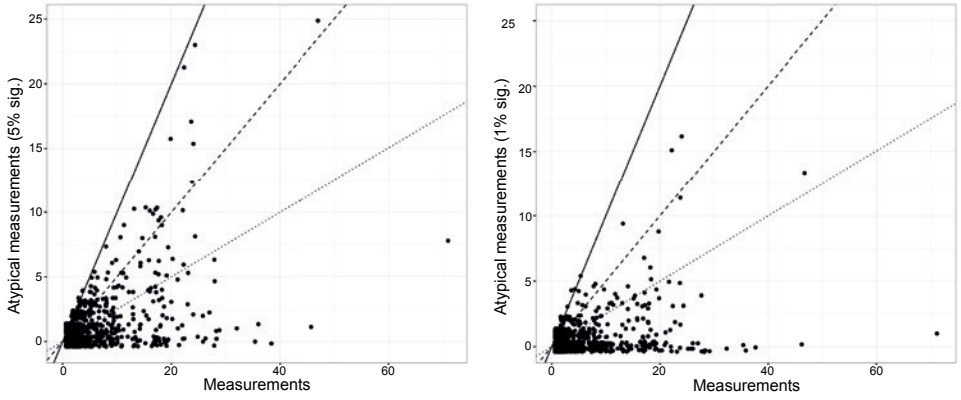
Please give an extensive explanation:

Other remarks regarding the inconsistency alert or the patient:

SUPPLEMENT 3. Differences in symptom profiles between the atypical and the non-atypical patterns as defined based on person-fit scores, with poor person-fit (l_z < -2.21) for atypical patterns and good person-fit for the non-atypical patterns (l_z > -2.21). Mean item scores are shown in the top panel with corresponding mean differences (positive scores represent more frequent presence in atypical patterns) in the bottom panel.



SUPPLEMENT 4. Relation between number of measurements and measurement flagged as atypical for each patient at 5% (left) and 1% (right) significance levels. The lines represent the point where all measurements are flagged as atypical (solid), 50% atypical (dashed), or 25% atypical (dotted).



CHAPTER 10

Discussion

DISCUSSION

The overall aim of the current thesis was to investigate how a data-analytical approach focused on disentangling the relation between symptoms and depression could (a) aid our understanding of depression heterogeneity, (b) be used to assess psychometric properties and improve depression measurement, and (c) be used to personalize depression assessments and provide additional individual feedback in clinical care.

The findings will be discussed in light of these three different perspectives, corresponding with the outline in the introduction. After a short summary of the main findings, the results of the different chapters will be integrated in a broader discussion from each perspective, touching on advances in current literature, limitations, and suggestions for further research.

SUMMARY OF MAIN FINDINGS

Many symptoms assessed for depression are non-specific for depression, causing differences in symptom patterns that do not reflect depression severity (**chapters 2 and 3**). These differences are also present across clinically relevant groups, and more accurate estimates of depression severity may be obtained by adjusting scores for these differences (**chapter 4**).

The analyses revealed some fundamental issues, which were further discussed and investigated in this thesis, including the differential weighting of individual symptoms (**chapter 5**) and the strong dependencies between symptoms that can be problematic when the aim is to identify homogenous subtypes (**chapter 6**).

In two large population studies (NEMESIS, n=5,583; Lifelines, n=73,403) results showed the usefulness of more advanced subtyping approaches to identify cross-diagnostic subtypes of depression and anxiety while incorporating measures of disability (**chapters 7 and 8**).

Finally, the clinical use of data-driven statistics was demonstrated in a pilot study on automated person-fit feedback in depression assessment informing clinicians on possible inconsistent reported symptom patterns in a clinical care setting (**chapter 9**).

UNDERLYING STRUCTURE OF DEPRESSIVE SYMPTOMS

In this thesis several approaches were used to investigate the underlying structure of depressive symptoms. The used IRT framework provided different tools to analyze the relation between symptoms and the assumed underlying depression severity. Differences and similarities across clinical manifest groups were investigated by means of DIF, as well as across latent groups by means of mixture models and person-fit statistics.

In the current section, the findings from the different chapters and analyses will be discussed in terms of (a) how they inform us about the complex multidimensionality that is inherent to psychiatric constructs like depression, (b) how data-driven research shows us that we should widen our focus and incorporate more symptoms and sources of relevant variability on top of those that are included in current classifications, and (c) whether we really need complex data-driven models to advance our understanding of depression.

MULTIDIMENSIONALITY WITHIN, BETWEEN, AND ACROSS SYMPTOMS

One important source of heterogeneity in depression is a range of secondary factors that cause symptoms of depression to be experienced for other reasons than the presence of MDD. This multifactorial nature of the heterogeneity of depressive symptomatology gives rise to complex multidimensionality across symptoms, between pairs of symptoms and within single symptoms.

Across symptoms secondary factors might lead to symptom patterns where for part of the pattern a different factor causes an increased or decreased likelihood of experiencing particular symptoms. There is probably an endless list of factors that could play such a role. The most obvious factor is perhaps the presence of a comorbid (or other primary) psychiatric or somatic disorder. For example, patients with an anxiety disorder (**chapters 2 and 4**), cardiovascular disease[1] (**chapter 3**), or chronic medical conditions[2] (**chapter 4**) are more likely to endorse a certain set of depressive symptoms. Medication may also be an under recognized secondary factor, where side effects can cause certain depressive symptoms[3] and treatment effects can differ across symptoms[4]. For instance, antidepressants may primarily elevate mood-related symptoms, with improvement of other symptoms as a secondary consequence that might take longer to occur. In addition, isolated symptoms may be present as residual symptoms when MDD is in remission[5], as a result of different clinical subtypes of depression[6], or can even be reported for secondary gains[7].

The presence of multidimensionality can also lead to pairs of symptoms being inherently dependent on each other. That is, the experience of one symptom directly tells us something about an increased chance of experiencing another symptom, unconditional on depression severity. When looking at depressive symptoms, several of such dependencies are thinkable, with some being connected stronger than others. For example, it is very intuitive to expect that for someone that experiences sleep disturbances it is also more likely to report fatigue, regardless of depression severity or the exact underlying causes. Such dependencies between symptoms are most likely not informative with regard to a person's depression. Therefore, these dependencies should ideally be accounted for, especially in analyses that could otherwise yield artificial results (**chapter 6**). To date, there has not been much research on the conditional dependencies between symptoms in psychiatric disorders like depression, but the increased attention to the analysis of individual symptoms[3] and the use of techniques like network analysis[8] and causal modeling[9] will probably shed much more light on this over the next decade.

Multidimensionality is also present within symptoms, where the same symptom can reflect something else in the context of different disorders. For example, concentration problems are considered a symptom in both the DSM-5 diagnostic criteria of MDD and the criteria of GAD. However, the type of experienced cognitive impairments has been observed to differ across depression and anxiety disorders[10]. In depression, the key cognitive impairment seems to be in the executive domain[11], whereas in anxiety disorders cognitive impairment seems to be more pronounced in attention and verbal memory[12], and studies report no executive dysfunction in anxiety disorders[10,13] or only in the presence of comorbid MDD[14]. Irrespective of the exact way in which specific cognitive problems are linked to specific disorders, the current use of very heterogeneous items (i.e. 'concentration problems' can mean different things) to assess cognitive problems in depression assessment makes it hard to find out what exactly is going on in a patient. The consequences of item heterogeneity for our understanding and measurement of the depression construct are discussed later in more detail from a psychometric perspective.

CASTING WIDER NETS ON DEPRESSION AND ANXIETY

Central to this thesis was an empirical approach to the study of differences and similarities in symptom patterns across individuals. Gaining this understanding is important because the heterogeneity of depression is thought to hamper scientific advances[15]. By finding groups of individuals that are more similar in terms of their symptom patterns[16] we might obtain homogenous phenotypes that are more useful to find specific biological[17] and neurological associations[18] and that could ultimately improve treatment and clinical care[19].

In this thesis, a data-driven approach that investigated a categorical, dimensional, and a mixture of both approaches revealed interesting insights in the phenotypical variations of depression and anxiety. First, the results showed that there are only naturally occurring subgroups with mixed symptomatology of both depression and anxiety. Second, the findings revealed that other sources of clinical variation such as measures of disability were found to be highly informative to optimally explain individual differences and subtyping. Third, it was observed that a large part of the population experiences symptoms of depression and anxiety at subthreshold levels, yet with associated disability. These findings were discussed in specific detail in **chapters 7 and 8**, of which the implications will be discussed here in a broader context.

An important observation in both studied population cohorts was that there was no clear empirical distinction between those who experience depression symptoms and those who experience anxiety symptoms. Instead, data-driven analyses revealed only mixed groups and no pure disorders. This is in line with a growing body of evidence for the strong etiological and phenomenological overlap between depression and anxiety, showing that both disorders load on a single internalizing dimension[20–23], have a strong shared heritability[24] and have shared environmental risk factors[25]. In addition, the disorders respond similarly to antidepressant treatment[26,27], psychosocial treatments[28] and self-guided help[29]. Interestingly, strong associations were also observed between the currently identified homogeneous subgroups and somatic/cardiovascular symptoms that have been suggested to play a dynamic role in depression and anxiety[30]. This raises the question whether, if the aim is to better understand the symptom-heterogeneity among persons, the nets should be cast even wider. This idea has motivated ambitious projects like HowNutsAreTheDutch[31], a crowd-sourcing study that explicitly takes a broad approach and assesses a wide range of mental complaints, but also many mental strengths. Studies like these might enable a different approach to the problem and might allow investigations of where the boundaries of the internalizing spectrum lie that both separate it from normality, and distinguishes it from different disorders[32]. One challenge to researchers seeking to broaden their focus in this way, is to lose the structure of psychopathology that is now often (consciously or unconsciously) ingrained in the obtained data, the used questionnaires, the applied models, and present in the minds of patients, test takers, and researchers. Otherwise, studies of the deeper structure of psychopathology will continue to yield results that mainly reflect the a priori structures imposed on the data.

Besides a broader set of symptoms, additional sources of clinically relevant variability should be considered when trying to understand interpersonal differences. Both in **chapters 7 and 8**, disability measures were incorporated in the developed subtyping models. This

was done because the presence of disability is a defining criterion for the pathological nature of symptoms and signals a need for care. In addition, disability levels are a source of clinically relevant variability and known to predict important clinical outcomes such as treatment response[33], remission[34,35], and recurrence[36]. Importantly, such relevant sources of variability should be included as covariates in the actual modeling process and not only as external variables later on to 'validate' previously estimated models. When aiming to optimally classify persons into clinically relevant subgroups, it makes no sense to keep relevant information from the used model (e.g. LCA, MM-IRT) and then later on hope that significant associations between the subgroups and the relevant information will be found to confirm the validity/usefulness of the model. Rather, one should endeavor to increase the amount of relevant known information in a measurement model as much as possible to enable optimal estimation of the unknown information that we are actually interested in.

Interestingly, the results reported in this thesis align with the previous finding that a large part of the population experiences symptoms of depression and anxiety without meeting the diagnostic criteria of a full blown disorder[37,38]. These 'Subclinical' individuals do however experience impaired functioning in the social and physical domains and report problems with work or other regular daily activities due to their mental problems (**chapter 8**). These findings furthermore show that the supposed dichotomy between normal and disordered states is empirically not supported[39]. The observed subclinical class may be a promising target population, in which to study the development of disorders, and to gain understanding of why some develop a full clinical disorder whereas others do not. The clinical implications of the fact that a large proportion of the population does experience subthreshold symptomatology and suggestions for how this knowledge could be utilized to improve clinical care are discussed later from a clinical perspective.

A SMOKESCREEN SURROUNDING COMPLEX MODELS?

Depression is a complex disorder with a complex etiology[40], characterized by complex interdependent relations between genetic, biological, environmental, and phenotypic variations[41]. Indeed, the results of this thesis fit in with this line of thought and unequivocal show that this multifactorial and complex nature is reflected in self-reported symptom-level data, where many factors seemingly play a role in determining whether someone experiences a symptom or not. The inherent complexities in the structure of depression and anxiety are often called upon to justify the use of more complex modelling approaches. These approaches bring with them a lot of analytical complexity and technical challenges that could by some be considered a smokescreen obscuring the true added value of these complex models. Is there really a need for more complex modelling?

On the one hand, the presence of so many factors that are known to influence depression makes it favorable to use more complex techniques. Modeling a complex disorder might simply ask for a more complex model. As argued in **chapter 6** with regard to the symptoms of appetite and weight change, there might be strong secondary factors underlying symptoms that are not informative about depression but that can dominate a model's results. With certain approaches that do not properly account for such effects, complexities in the data may lead to biased or artificial results. We therefore suggest the use of more advanced methods that are capable of modelling both the wanted and unwanted effects, allowing to study the true effect of interest. Although statistically more complex, these models can actually yield a solution that is easier to interpret. For instance, in **chapters 7 and 8** the most complex MM-IRT-C models led to solutions with fewer and easier to interpret classes than the less complex LCA models.

On the other hand, complex models run the risk of modeling all kinds of irrelevant or meaningless information, especially in large datasets that contain so much information that there is always something extra to be modeled. Increased complexity always allows for a better description of the data, but this comes with the risk of overfitting and results that generalize poorly. Selecting the best model given a dataset is a whole topic on its own in statistical research, and statistics, such as information criteria are used to penalize a model's increased fit for its increased complexity[42]. Also, the risk of overfitting can be reduced by utilizing techniques like bootstrapping[43,44] or cross-validation[45,46]. Different criteria can however point at different models, and it is often unclear if they really retrieve the model that best reflects the construct as it is in reality. Which model selection criteria to use is one of many choices a researcher has to make, but also include choices in pre-modelling procedures like variable selection and data imputation. In addition, there are many subjective choices that need to be made during the actual modeling as there are many options when it comes to selecting an estimation procedure and model definitions (e.g. restrictions on parameters). Different subjective choices can lead to a variety of alternative models[47,48] making it more difficult to assess how good the selected model is. In many cases, it is safe to assume that when results from a single complex model are presented, dozens of models have been estimated in reality.

Many studies that use more advanced complex models are focused exclusively on selecting the best-fitting model. The truth is that in many cases the discarded models can still be a good description of most of the data. That is, if selection criteria point at a 5-class model as optimal than this does not mean that the 4-class model (or the 6-class model) is complete nonsense. In most cases, the discarded models are likely to be perfectly capable of modeling most patterns in the data. In latent variable modeling there may be several other ways to investigate if and why one model could outperform another.

First, results of alternative models should be investigated (i.e. with fewer or more classes, or for example continuous versus categorical approaches) and not only results from the selected model. In **chapter 8**, the results clearly indicated that LCA models were suboptimal when compared to the mixture-IRT models. However, it was still very informative to have looked at all the different LCA solutions as it revealed the presence of possible categorical and dimensional effects in the data. When using suboptimal models in this way, it is relatively unimportant that the models do not describe things how they really are, because they can still serve as useful tools to better understand and describe the data.

Second, there are promising developments to compare models and understand them in more detail. The potential influence of secondary factors on the selected model can be assessed by modeling latent secondary dimensions[49] or by investigating their expected influence on model parameters[50] without the need of exploring all possible alternative models. Furthermore, models can be evaluated on the person-level by looking at the individual influences on model selection[51] or by investigating each individual's distance to the model, which can be used to evaluate how well the model fits for everyone[52]. Such analyses might help to gain better insight into what more complex models of depression heterogeneity are actually modeling and to decide whether this thing is meaningful or not. For example, Reise and colleagues[52] show that although a bi-factor model might obtain better fit for a given sample, a more simple unidimensional model can fit perfectly well for most people (>85%). In this case, the increased complexity of the bi-factor model is needed only to account for the response patterns of a minority of the sample.

In summary, there lies danger in the complexity of depression data where models can be modeling all kinds of effects that are not relevant to the construct itself. To lift the smokescreen surrounding complex models we need to obtain a better understanding of why one model describes the data better than alternative models, instead of obtaining a number that reflects it does.

THE MEASUREMENT AND MISMEASUREMENT OF DEPRESSION

The study of depression heterogeneity is necessarily interwoven with the measurement of depression. It is vital to have an accurate and valid measurement of the depression phenotype if we are to understand individual differences and to find relevant associations with genetic, biological, and environmental factors[41]. In this thesis, several approaches were investigated that could be used to achieve increased measurement precision and inform on the validity of depression assessments.

IMPROVING PRECISION OF MEASURING THE DEPRESSION PHENOTYPE

The findings in this thesis showed that measurement precision can be increased by adjusting scores for DIF (**chapter 4**), or by obtaining more appropriate, weighted item scores (**chapter 5**). In **chapter 4** we showed that raw total scores were robust against the presence of DIF, limiting the potential clinical impact of item bias across groups. This robustness is mostly explained by the observation that correlations between raw severity scores and optimized IRT-based severity scores always remain very high (often with r>0.9). This high correlation results from the fact that only little changes in the overall severity rank order of patients when using IRT-based scores instead of raw scores: patients scoring high on depression severity in raw total scores will still score high when optimal scores are used, and vice versa. The robustness of severity scores to DIF is reassuring for those using depression scores in clinical depression assessment and monitoring. However, researchers that study depression may still want to use optimal IRT scores that offer a more precise measurement of the phenotype and are more sensitive to change in depression severity. Studies have shown that using optimal scores less often leads to finding spurious interactions[53] and leads to more reliable conclusions in gene-environment studies[54].

Another important aspect of increased measurement precision is having the right measurement model for the right person. Person-fit statistics can provide insight into this, allowing for the identification of those that report inconsistently under a current measurement model. For these identified persons, the estimated total scores are not a valid reflection of depression severity and may be an over- or underestimation. This information can be used to increase measurement precision in two ways. First, it can be used in an iterative estimation (e.g. IRLS[55]) where the patterns that poorly fit the model get down-weighted during estimation, leading to more purified estimates of the measurement model. Second, different measurement models may be explored in the 'misfitting' group. For example, person-fit statistics were used in **chapter 4** to investigate depressive symptom reporting in a group of heart patients. The model on the group level was characterized by a dominant role for somatic symptoms. As a result, many depressed patients were allocated to the 'misfitting' group. Refitting a model in this latter group could reveal a measurement model with a less dominant role for somatic symptoms, which better describes the response patterns of truly depressed heart patients. As such, person-fit can be used as a tool to explore if different measurement models may hold in a sample, ultimately allowing for the improvement of measurement precision in those individuals identified as misfitting.

ITEM AND CATEGORY FUNCTIONING

In **chapters 2, 3, and 4** several items had to be removed and categories had to be recoded to achieve the item quality needed for the purpose of analyses. Using nonparametric IRT, item quality and assumptions of the performed analyses were assessed, which led to removal of several items because they violated model assumptions and could bias the results. In **chapter 5**, we followed up on these results with a more detailed analysis on item- and category functioning in a depression questionnaire. Results showed that for some items, category functioning was poor, with unordered and redundant categories. In addition, some items were shown to have little discriminatory power across increasing response categories, indicating dichotomous (symptom present or not) rather than ordinal item functioning. Taken together, these results showed that there are serious fundamental validity issues underlying items that measure individual depressive symptoms. This leads to fundamental questions underlying the measurement of depression. Can all symptoms of depression be meaningfully measured along a dimension of symptom severity? Are depressive symptoms interchangeable, as assumed in the current diagnostic systems? Especially when items are used in symptom-level analyses, closer inspection is needed first to assess whether single items validly measure actual symptom severity.

HETEROGENEOUS INSTRUMENTS FOR A HETEROGENEOUS CONSTRUCT

Psychometrically sound instruments have been developed for the measurement of psychiatric constructs like depression in, for example, projects like the Patient Reported Outcomes and Measurement Information Systems (PROMIS[56]), and instruments like the Inventory of Depression and Anxiety Symptoms (IDAS[57]). In these projects, state-of-the-art methods, such as factor analysis[57] and IRT methods[56], were used to obtain psychometrically sound instruments. However, in the construction of all these questionnaires, there is a strong guiding assumption that good scales need to be homogenous and unidimensional. There are several aspects to psychiatric constructs that perhaps warrant a different point of view. Aiming for homogeneity and unidimensionality could lead to less useful scales[58] that give a limited representation of the depression construct[59]. Indeed, many psychometric scales that are nicely homogenous measure a smaller construct where they especially lack measurement information at the more severe end of the spectrum[60]. It is more logical to instead measure the heterogeneous construct of depression by a heterogeneous set of items that capture all necessary variation and together provide a good and complete reflection of depression severity. *"The meaningfulness of a test lies not in a methodological prescription of homogeneity but in the test's ability to capture all relevant attributes of the entity it purports to measure"*[61]. Two additional, more technical arguments can also be made in favor of a more heterogeneous friendly perspective on measuring depression.

First, many factors may drive the response on an item. There are nearly no perfect items for depression (besides perhaps the core symptoms of sad mood and loss of interest) that are truly unidimensional and only reflect depression. There is always some contamination by multidimensionality, with symptoms being reported for other reasons than depression. For example, results in **chapter 8** show that a large part of the population reports sleep disturbances and fatigue without reporting any other symptoms, suggesting that these symptoms are not reflective of depression. Sleep disturbances are seen in the general population due to, for example, work stress[62], or due to medical conditions[63,64]. However they are also part of depression, and part of depression questionnaires. It is important to realize that these symptoms reflect depression severity when in the presence of other depressive symptoms, and that despite the item heterogeneity they are very relevant to the depression construct.

Second, an item can be informative only for part of the depression severity scale. These items show so-called pseudo-trait like behavior, where at a large part of the depression severity scale the responses are not reflective of depression severity. This might especially be true for more extreme symptoms like 'weight change' and 'thoughts of death' that have previously been suggested as symptoms that could be removed to obtain more homogeneous depression scales[65]. A considerable part of the individuals reporting these symptoms will do so for other reasons than depression, because it is not uncommon for many individuals, depressed or not, to have a change of weight or to think about death (e.g. when confronted with a death in the family[66,67]). However, this does not mean that these symptoms have no role to play in depression measurement; at the severe end of the spectrum these symptoms start playing a more important role, and both weight change[66] and thoughts of death[68] are associated with increased depression severity and poor prognosis. As such, although non-informative in many respondents at the less severe side of the depression scale, these symptoms may serve as very useful indicators to discriminate between patients at more severe levels of depression.

There are some studies that have suggested alternative, non-standard models that might better describe the measurement of psychiatric constructs like depression. These include models for non-normal data[69], models that allow for unidimensional scaling in the presence of multidimensionality[70], the four-parameter model[71], and the unipolar trait model[72]. Of these, the latter two will be shortly discussed below.

It has been suggested that a four-parameter IRT model that allows for both lower and higher asymptotes would be more appropriate for use in psychopathology[71,73]. In these models, a lower asymptote can account for a baseline probability (prevalence) where individuals that are not depressed still have a probability of reporting the symptom, while a higher asymptote accounts for the fact that even the most extremely depressed patients

still do not have a near to 100% probability of reporting the symptom. The authors showed that many items in psychopathology require to set lower and higher asymptotes that when estimated are informative on item characteristics, and more appropriately reflects the response process underlying psychopathology items.

The unipolar trait model works from the assumption that not everyone can be meaningfully scaled along the continuum. Standard IRT models assume a bipolar trait, where going up the scale reflects increasing severity to infinity, and going down the scale reflects decreasing severity to infinity. Here, it is assumed that meaningful scaling is also possible for those individuals that are non-depressed in terms of depression severity, and that it is still possible to say for two non-depressed individuals that one non-depressed individual is less non-depressed than the other. In a unipolar trait model this paradox is prevented by anchoring the scale at the lowest level for which depression severity can be meaningfully assessed. All measurements of depression severity would then be relative to this anchor. In this approach the lower part of the traditional scale is thus not assumed to measure the construct, making it more suitable to model items that show pseudo-trait behavior (see discussion above). The unipolar trait model was shown to be useful in modeling addictive disorders (Lucke, 2014), where the model behaves more in line with theories emphasizing that worsening of the disorder should be assessed relative to the baseline level of no disorder.

DATA-DRIVEN DECISION SUPPORT

The effectiveness and efficiency of depression care may be improved by tailoring interventions and support to individual needs. Data-analytical tools may play an important role in helping to achieve this goal. In several chapters of this thesis, the hypothesized contribution of data-driven techniques to improve clinical care was implied and discussed. This led to a first pilot study on the potential clinical usefulness of individual feedback based on person-fit statistics within depression care (**chapter 9**). Below, it is discussed how personalizing depression assessments and providing automated feedback in case of inconsistent responses could lead to improvements in clinical care.

In **chapter 9**, a pilot implementation study was described that was the first to investigate if and how a person-fit statistic could be used to alert clinicians about possible inconsistent reporting of depressive symptoms in clinical depression care. The findings showed that, in retrospect, psychiatrists judged such an alert of potential clinical use, suggesting that it could lead to new insights, be a reason for discussion with the patient, and help the interpretation of depression assessments that do not conform to the clinical expectation. Furthermore, the pilot showed that psychiatrists were perfectly capable of assessing

the value of an alert that is based completely on statistical information and putting the information into a clinical context. Clinical decision-making is a process of collecting and summarizing information that is translated into guidance for treatment[74,75]. In this process, data-driven approaches, such as person-fit statistics, could provide an additional source of information that is easy to collect and of clinical use. Importantly, such systems should not be seen as a replacement for a clinician's tasks within the decision-making process, but rather as a supportive tool that gives extra diagnostic information in the form of personalized feedback that summarizes essential information in the available data.

Apart from the implementation of above described person-fit based alerts, there are also other ways, in which the IRT methods used in this thesis allow for a more personalized approach to depression assessment. In **chapter 4**, we used IRT to obtain estimates of depression severity that were adjusted for differential item functioning across clinical groups, tailoring the measurement to individual characteristics. In **chapter 9,** IRT was used to identify patients with atypical symptom patterns for which other factors besides depression might play an important role, and who therefore need a different approach in depression assessment. These research examples show how we could improve applied measurement precision by utilizing information from advanced models. New developments also include computerized adaptive testing (CAT[76]) where the most optimal set of items is adaptively selected after each given response from a large pool of items to obtain a more precise measurement. Such approaches could be very promising in the future of depression assessment and monitoring, yielding a more detailed and precise measurement that might be crucial to tailoring treatment and support to individual needs[77,78].

The results in this thesis also provide clues about particular groups that could benefit from improved measurement and interventions. Recent figures from the Netherlands showed that 52% of patients, who are seen by the general practitioner with mental complaints end up on a waiting list[79]. This time can already be used to collect information on the complaints of the patient, and gives an opportunity for data-driven approaches to summarize information that may be used to stratify patients for certain interventions. In **chapter 7**, it was suggested that the differences between a 'Subclinical' and 'Clinical' class could be utilized to develop such an early stratification tool. That is, those that fit the symptom patterns and experienced disability associated with the subclinical class could be identified and offered low-cost, effective primary care treatment programs[80], self-help treatments[29] or guided online treatment programs[81]. Those with more severe clinical psychopathology could be referred to specialized care. Such a system would allow to tailor interventions, but also to offer care to those that otherwise might not be judged as having a clinical disorder but that still experience disability and could benefit from low-threshold care.

CONCLUDING REMARKS

This thesis showed the value of data-driven analytical tools in studying the complex structure of depression, and its potential to improve depression assessment and care. Findings advocate a shift in the focus of research towards individual depressive symptoms, and towards closer evaluation and comparison of different complex models, where the structure of depression is more complex than can be addressed by a single statistical technique. Furthermore, ample empirical evidence was found for many aspects of current diagnostic systems. Diagnostic nets should be cast wider in research and clinical practice to allow for a clearer focus of efforts to decrease the burden and costs of depression and anxiety in the population.

The framework of IRT has much to offer for psychiatry, besides improving measurement precision that can give researchers stronger phenotypes, it has the potential for clinicians to improve care with personalized assessments, and individualized feedback. Implementation studies might be key in bridging a gap between clinicians and researchers, leading to advances in both fields, where these studies may reveal the obstacles that we need to overcome in research, and at the same time show the potential offered by data-driven decision support to clinicians that ultimately lead to improved mental health care.

REFERENCES

1. Hoen, P. *et al.* Differential associations between specific depressive symptoms and cardiovascular prognosis in patients with stable coronary heart disease., Differential associations between specific depressive symptoms and cardiovascular prognosis in patients with stable coronary heart disease. *J. Am. Coll. Cardiol. J. Am. Coll. Cardiol.* **56, 56,** 838, 838–844 (2010).
2. Yang, F. M. & Jones, R. N. Measurement Differences in Depression: Chronic Health-Related and Sociodemographic Effects in Older Americans. *Psychosom. Med.* **70,** 993–1004 (2008).
3. Fried, E. I. & Nesse, R. M. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med.* **13,** 72 (2015).
4. Harmer, C. J., Goodwin, G. M. & Cowen, P. J. Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *Br. J. Psychiatry* **195,** 102–108 (2009).
5. Conradi, H. J., Ormel, J. & Jonge, P. de. Presence of individual (residual) symptoms during depressive episodes and periods of remission: a 3-year prospective study. *Psychol. Med.* **41,** 1165–1174 (2011).
6. Lichtenberg, P. & Belmaker, R. H. Subtyping Major Depressive Disorder. *Psychother. Psychosom.* **79,** 131–135 (2010).
7. Rogers, R., Sewell, K. W., Martin, M. A. & Vitacco, M. J. Detection of feigned mental disorders: a meta-analysis of the MMPI-2 and malingering. *Assessment* **10,** 160–177 (2003).
8. Borsboom, D. & Cramer, A. O. J. Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annu. Rev. Clin. Psychol.* **9,** 91–121 (2013).
9. Kinderman, P. *et al.* Causal and mediating factors for anxiety, depression and well-being. *Br. J. Psychiatry* **206,** 456–460 (2015).
10. Castaneda, A. E., Tuulio-Henriksson, A., Marttunen, M., Suvisaari, J. & Lönnqvist, J. A review on cognitive impairments in depressive and anxiety disorders with a focus on young adults. *J. Affect. Disord.* **106,** 1–27 (2008).
11. Stordal, K. I. *et al.* Impairment across executive functions in recurrent major depression. *Nord. J. Psychiatry* **58,** 41–47 (2004).
12. Airaksinen, E., Larsson, M. & Forsell, Y. Neuropsychological functions in anxiety disorders in population-based samples: evidence of episodic memory dysfunction. *J. Psychiatr. Res.* **39,** 207–214 (2005).
13. Boldrini, M. *et al.* Selective cognitive deficits in obsessive-compulsive disorder compared to panic disorder with agoraphobia. *Acta Psychiatr. Scand.* **111,** 150–158 (2005).
14. Kaplan, J. S. *et al.* Differential performance on tasks of affective processing and decision-making in patients with Panic Disorder and Panic Disorder with comorbid Major Depressive Disorder. *J. Affect. Disord.* **95,** 165–171 (2006).
15. Baumeister, H. & Parker, G. Meta-review of depressive subtyping models. *J. Affect. Disord.* **139,** 126–140 (2012).
16. van Loo, H. M., de Jonge, P., Romeijn, J.-W., Kessler, R. C. & Schoevers, R. A. Data-driven subtypes of major depressive disorder: a systematic review. *BMC Med.* **10,** 156 (2012).
17. Kapur, S., Phillips, A. G. & Insel, T. R. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* **17,** 1174–1179 (2012).
18. Cuthbert, B. N. & Insel, T. R. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* **11,** 126 (2013).
19. Andrews, G., Anderson, T. m., Slade, T. & Sunderland, M. Classification of Anxiety and Depressive disorders: problems and solutions. *Depress. Anxiety* **25,** 274–281 (2008).
20. Krueger, R. F. & Bezdjian, S. Enhancing research and treatment of mental disorders with dimensional concepts: toward DSM-V and ICD-11. *World Psychiatry* **8,** 3–6 (2009).
21. Eaton, N. R. *et al.* The structure and predictive validity of the internalizing disorders. *J. Abnorm. Psychol.* **122,** 86–92 (2013).
22. Kushner, M. G. *et al.* Modeling and treating internalizing psychopathology in a clinical trial: a latent variable structural equation modeling approach. *Psychol. Med.* **43,** 1611–1623 (2013).
23. Wright, A. G. C. & Simms, L. J. A metastructural model of mental disorders and pathological personality traits. *Psychol. Med.* **45,** 2309–2319 (2015).

24. Kendler, K. S. Genetic Epidemiology in Psychiatry: Taking Both Genes and Environment Seriously. *Arch. Gen. Psychiatry* **52,** 895–899 (1995).
25. Kendler, K. S., Prescott, C. A., Myers, J. & Neale, M. C. The Structure of Genetic and Environmental Risk Factors for Common Psychiatric and Substance Use Disorders in Men and Women. *Arch. Gen. Psychiatry* **60,** 929–937 (2003).
26. Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A. & Rosenthal, R. Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. *N. Engl. J. Med.* **358,** 252–260 (2008).
27. Roest, A. M. *et al.* Reporting Bias in Clinical Trials Investigating the Efficacy of Second-Generation Antidepressants in the Treatment of Anxiety Disorders: A Report of 2 Meta-analyses. *JAMA Psychiatry* **72,** 500–510 (2015).
28. Haby, M. M., Donnelly, M., Corry, J. & Vos, T. Cognitive behavioural therapy for depression, panic disorder and generalized anxiety disorder: a meta-regression of factors that may predict outcome. *Aust. N. Z. J. Psychiatry* **40,** 9–19 (2006).
29. Cuijpers, P., Donker, T., Straten, A. van, Li, J. & Andersson, G. Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. *Psychol. Med.* **40,** 1943–1957 (2010).
30. Bekhuis, E., Boschloo, L., Rosmalen, J. G. M. & Schoevers, R. A. Differential associations of specific depressive and anxiety disorders with somatic symptoms. *J. Psychosom. Res.* **78,** 116–122 (2015).
31. Krieke, L. V. D. *et al.* HowNutsAreTheDutch (HoeGekIsNL): A crowdsourcing study of mental symptoms and strengths. *Int. J. Methods Psychiatr. Res.* **25,** 123–144 (2016).
32. Ruscio, J. & Ruscio, A. M. Clarifying Boundary Issues in Psychopathology: The Role of Taxometrics in a Comprehensive Program of Structural Research. *J. Abnorm. Psychol.* **113,** 24–38 (2004).
33. Hirschfeld, R. M. A. *et al.* Does psychosocial functioning improve independent of depressive symptoms? a comparison of nefazodone, psychotherapy, and their combination. *Biol. Psychiatry* **51,** 123–133 (2002).
34. Von Korff, M. *et al.* Effect on disability outcomes of a depression relapse prevention program. *Psychosom. Med.* **65,** 938–943 (2003).
35. Zimmerman, M. *et al.* Remission in depressed outpatients: More than just symptom resolution? *J. Psychiatr. Res.* **42,** 797–801 (2008).
36. Scholten, W. D. *et al.* Recurrence of anxiety disorders and its predictors. *J. Affect. Disord.* **147,** 180–185 (2013).
37. Das-Munshi, J. *et al.* Public health significance of mixed anxiety and depression: beyond current classification. *Br. J. Psychiatry* **192,** 171–177 (2008).
38. Hettema, J. M., Aggen, S. H., Kubarych, T. S., Neale, M. C. & Kendler, K. S. Identification and validation of mixed anxiety–depression. *Psychol. Med.* **45,** 3075–3084 (2015).
39. Kendler, K. S. The dappled nature of causes of psychiatric illness: replacing the organic–functional/hardware–software dichotomy with empirically based pluralism. *Mol. Psychiatry* **17,** 377–388 (2012).
40. Kendler, K. S., Gardner, C. O. & Prescott, C. A. Toward a Comprehensive Developmental Model for Major Depression in Women. *Am. J. Psychiatry* **159,** 1133–1145 (2002).
41. Craddock, N. & Owen, M. J. The Kraepelinian dichotomy – going, going... but still not gone. *Br. J. Psychiatry* **196,** 92–95 (2010).
42. Lubke, G. & Muthén, B. O. Performance of Factor Mixture Models as a Function of Model Size, Covariate Effects, and Class-Specific Parameters. *Struct. Equ. Model. Multidiscip. J.* **14,** 26–47 (2007).
43. Mclachlan, G. On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Appl. Stat.* **36,** 318–324 (1987).
44. Nylund, K. L., Asparouhov, T. & Muthén, B. O. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Struct. Equ. Model. Multidiscip. J.* **14,** 535–569 (2007).
45. Kohavi, R. Kohavi, Ron. 'A study of cross-validation and bootstrap for accuracy estimation and model selection.' Ijcai. Vol. 14. No. 2. 1995. *Ijcai* **14,** 1137–1145 (1995).
46. Vermunt, J. & Magidson, J. *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. (Statistical Innovations Inc., 2013).

47. Beauchaine, T. P. & Waters, E. Pseudotaxonicity in MAMBAC and MAXCOV analyses of rating-scale data: Turning continua into classes by manipulating observer's expectations. *Psychol. Methods* **8,** 3–15 (2003).
48. Lubke, G. H. & Miller, P. J. Does nature have joints worth carving? A discussion of taxometrics, model-based clustering and latent variable mixture modeling. *Psychol. Med.* **45,** 705–715 (2015).
49. De Boeck, P. *et al*. Explanatory Secondary Dimension Modeling of Latent Differential Item Functioning. *Appl. Psychol. Meas.* **35,** 583–603 (2011).
50. Oberski, D. L., Vermunt, J. K. & Moors, G. B. D. Evaluating Measurement Invariance in Categorical Data Latent Variable Models with the EPC-Interest. *Polit. Anal.* **23,** 550–563 (2015).
51. Sterba, S. K. & Pek, J. Individual influence on model selection. *Psychol. Methods* **17,** 582–599 (2012).
52. Reise, S. P., Kim, D. S., Mansolf, M. & Widaman, K. F. Is the Bifactor Model a Better Model or Is It Just Better at Modeling Implausible Responses? Application of Iteratively Reweighted Least Squares to the Rosenberg Self-Esteem Scale. *Multivar. Behav. Res.* **51,** 818–838 (2016).
53. Kang, S.-M. & Waller, N. G. Moderated Multiple Regression, Spurious Interaction Effects, and IRT. *Appl. Psychol. Meas.* **29,** 87–105 (2005).
54. Murray, A. L., Booth, T. & Molenaar, D. When Middle Really Means 'Top' or 'Bottom': An Analysis of the 16PF5 Using Bock's Nominal Response Model. *J. Pers. Assess.* **98,** 319–331 (2016).
55. Yuan, K.-H. & Bentler, P. M. Robust mean and covariance structure analysis through iteratively reweighted least squares. *Psychometrika* **65,** 43–58 (2000).
56. Cella, D. *et al.* The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J. Clin. Epidemiol.* **63,** 1179–1194 (2010).
57. Watson, D. *et al.* Development and validation of the Inventory of Depression and Anxiety Symptoms (IDAS). *Psychol. Assess.* **19,** 253–268 (2007).
58. Gustafsson, J.-E. & Åberg-Bengtsson, L. in *Measuring psychological constructs: Advances in model-based approaches* 97–121 (American Psychological Association, 2010).
59. Some thoughts concerning the recent shift from measures with many items to measures with few items. *Eur. J. Psychol. Assess.* **27,** 71–72 (2011).
60. Wahl, I. *et al.* Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. *J. Clin. Epidemiol.* **67,** 73–86 (2014).
61. Lucke, J. F. The of α and the ω of Congeneric Test Theory: An Extension of Reliability and Internal Consistency to Heterogeneous Tests. *Appl. Psychol. Meas.* **29,** 65–81 (2005).
62. Ota, A. *et al.* Association between psychosocial job characteristics and insomnia: an investigation using two relevant job stress models—the demand-control-support (DCS) model and the effort-reward imbalance (ERI) model. *Sleep Med.* **6,** 353–358 (2005).
63. Ohayon, M. M. Epidemiology of insomnia: what we know and what we still need to learn. *Sleep Med. Rev.* **6,** 97–111 (2002).
64. Phillips, B. A. *et al.* Sleep Disorders and Medical Conditions in Women. *J. Womens Health* **17,** 1191–1199 (2008).
65. Zimmerman, M., McGlinchey, J. B., Young, D. & Chelminski, I. Diagnosing major depressive disorder I: A psychometric evaluation of the DSM-IV symptom criteria. *J. Nerv. Ment. Dis.* **194,** 158–163 (2006).
66. Heiskanen, T. H. *et al.* Depression and major weight gain: A 6-year prospective follow-up of outpatients. *Compr. Psychiatry* **54,** 599–604 (2013).
67. Bauer, A. M., Chan, Y.-F., Huang, H., Vannoy, S. & Unützer, J. Characteristics, Management, and Depression Outcomes of Primary Care Patients Who Endorse Thoughts of Death or Suicide on the PHQ-9. *J. Gen. Intern. Med.* **28,** 363–369 (2013).
68. Busch, K. A., Fawcett, J. & Jacobs, D. G. Clinical correlates of inpatient suicide. *J. Clin. Psychiatry* **64,** 14–19 (2003).
69. Molenaar, D., Dolan, C. V. & Boeck, P. de. The Heteroscedastic Graded Response Model with a Skewed Latent Trait: Testing Statistical and Substantive Hypotheses Related to Skewed Item Category Functions. *Psychometrika* **77,** 455–478 (2012).

70. Ip, E. & Chen, S. in *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (Routledge, 2014).
71. Reise, S. P. & Waller, N. G. How many IRT parameters does it take to model psychopathology items? *Psychol. Methods* **8,** 164–184 (2003).
72. Lucke, J. in *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (Routledge, 2014).
73. Waller, N. G. & Reise, S. P. in *Measuring psychological constructs: Advances in model-based approaches* 147–173 (American Psychological Association, 2010).
74. Puschner, B. *et al.* Clinical Decision Making and Outcome in Routine Care for People with Severe Mental Illness (CEDAR): Study protocol. *BMC Psychiatry* **10,** 90 (2010).
75. Scott, K. & Lewis, C. C. Using Measurement-Based Care to Enhance Any Treatment. *Cogn. Behav. Pract.* **22,** 49–59 (2015).
76. Gibbons, R. D. *et al.* Development of a Computerized Adaptive Test for Depression. *Arch. Gen. Psychiatry* **69,** 1104–1112 (2012).
77. Rush, A. J. Isn't It About Time to Employ Measurement-Based Care in Practice? *Am. J. Psychiatry* **172,** 934–936 (2015).
78. Bickman, L., Lyon, A. R. & Wolpert, M. Achieving Precision Mental Health through Effective Assessment, Monitoring, and Feedback Processes. *Adm. Policy Ment. Health* **43,** 271–276 (2016).
79. Landelijke Huisartsen Vereniging (LHV). Ggz-peiling 2016 statistieken. Retrieved June 27, 2016 from http://www.lhv.nl. (2016).
80. Patel, V. *et al.* Effectiveness of an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in Goa, India (MANAS): a cluster randomised controlled trial. *The Lancet* **376,** 2086–2095 (2010).
81. Kenter, R. M. F. *et al.* Effectiveness and cost effectiveness of guided online treatment for patients with major depressive disorder on a waiting list for psychotherapy: study protocol of a randomized controlled trial. *Trials* **14,** 412 (2013).

Nederlandse samenvatting

Dankwoord

Curriculum vitae

List of publications

NEDERLANDSE SAMENVATTING

Depressie is wereldwijd één van de meest voorkomende ziektes en tevens één van de grootste gezondheidsproblemen van deze tijd. Meer dan 1 uit 7 mensen krijgt een depressie, waarbij het voor de helft van de mensen om een terugkerende ziekte gaat. De gevolgen van depressie zijn vergaand en leidt tot beperkingen in functioneren op sociaal, lichamelijk en werk vlak. Voor veel mensen wordt geen effectieve behandeling gevonden, waarbij meer dan 60 procent van de beperkingen die de depressie met zich meebrengt intact blijft. Depressieve patiënten verschillen sterk van elkaar, zowel in de klachten die ze hebben als het verloop van hun depressie over tijd. Het onderzoek naar depressie vordert maar langzaam en we begrijpen nog maar weinig van alle verschillen tussen depressieve patiënten.

De huidige classificatie van depressie ligt onder druk. Waar het voor behandelaren een bruikbaar hulpmiddel is in de klinische praktijk, wordt de bruikbaarheid binnen onderzoek steeds vaker in twijfel getrokken. Stoornissen komen vaak samen voor en houden zich niet aan de grenzen van de diagnostische classificatie. Mensen met dezelfde stoornis kunnen zo sterk van elkaar verschillen dat het onaannemelijk is dat hier dezelfde oorzaken aan ten grondslag liggen. Dit bemoeilijkt onderzoek naar depressie en roept op tot het ontwikkelen van alternatieven voor huidige classificatiesystemen.

In dit proefschrift probeerden we met behulp van statistische modellen op een andere manier naar depressie te kijken. Hierbij namen we niet aan dat het huidige diagnostische systeem klopt en probeerden we verder te kijken dan de diagnose depressie alleen. Dit houdt onder andere in dat: (i) behalve depressieve symptomen ook andere symptomen mee werden genomen, zoals symptomen van angst, (ii) naast symptomen ook andere klinisch relevante indicatoren werden meegenomen, waaronder problemen in sociaal en lichamelijk functioneren, en (iii) onze analyses niet beperkt waren tot alleen mensen die voldoen aan de criteria voor een diagnose depressie. Door onze blik te verbreden en door het toepassen van een data-gedreven benadering, levert dit proefschrift op drie verschillende wijzen een bijdrage aan het huidige depressie onderzoek. Allereerst kunnen de resultaten ons meer vertellen over de patronen van symptomen waarmee het ons begrip van individuele verschillen binnen depressie vergroot. Daarnaast geven de gebruikte modellen inzicht in het meten van depressie en bieden ze mogelijkheden om dit met meer precisie te doen. Als laatste kunnen de data-gedreven modellen ingezet worden als bruikbare hulpmiddelen in de klinische praktijk, wat kan leiden tot effectievere zorg.

SYMPTOMEN EN DEPRESSIE

Een belangrijke oorzaak voor het gebrek aan doorbraken in depressie onderzoek wordt vaak geweten aan de grote verschillen tussen patiënten. Geen patiënt is hetzelfde: ze reageren anders op behandeling, ervaren andere symptomen waar elk symptoom op zichzelf ook weer een andere rol kan spelen, en laten een ander beloop over tijd zien. Data-gedreven modellen werden gebruikt om meer inzicht te krijgen in deze heterogeniteit.

In hoofdstuk 2 deden we dit door in de data een groep te identificeren met patronen van depressieve symptomen die inconsistent zijn met het te verwachten patroon. Door middel van een statistisch model werd eerst vastgesteld hoe de relatie tussen symptomen en depressie bij de meeste mensen eruitziet: dit geeft inzicht in de vraag welke symptomen mild van aard zijn (bijvoorbeeld piekeren) en welke symptomen juist ernstig zijn (bijvoorbeeld gedachten aan de dood). Vervolgens konden we voor elke persoon het patroon van gerapporteerde symptomen vergelijken met dit model en vaststellen hoe typisch of atypisch dit patroon is voor depressie. Deze vernieuwende benadering bleek succesvol en nuttig te zijn. Het liet ons zien dat er veel factoren zijn naast depressie die ervoor kunnen zorgen dat iemand een depressie symptoom ervaart. Dit heeft als gevolg dat veel verschillen tussen patiënten in de symptomen die zij ervaren niet reflectief zijn voor een echt verschil in de ernst van depressie.

In hoofdstuk 3 gebruikten we dezelfde methode in een groep van hartpatiënten om patiënten te identificeren met inconsistente symptoompatronen. De bevindingen lieten zien dat de somatische symptomen van depressie een te groot gewicht hebben in deze groep. Deze symptomen passen beter bij de directe gevolgen van de hartziekte en reflecteren in mindere mate de ernst van depressie. In hoofdstuk 4 laten we zien dat vergelijkbare factoren ook een bron van heterogeniteit zijn in bestaande klinisch relevante groepen. Zo bleken patiënten die even depressief zijn verschillende symptomen te rapporteren wanneer ze in eerstelijns of tweedelijns zorg zitten, wel of geen chronische ziekte hebben, en wel of geen angstdiagnose hebben.

De bevindingen in deze hoofdstukken riepen enkele fundamentele vragen op die bij het onderzoek naar depressie een belangrijke belemmering kunnen vormen. In hoofdstuk 5 werd nader gekeken naar het meten van individuele symptomen middels enkele items en de daarbij horende antwoord-categorieën. In huidig onderzoek en diagnostiek wordt bij het berekenen van ernstscores aan alle symptomen gelijk gewicht gegeven, waarbij het in principe niet uitmaakt welke symptomen je hebt naast de twee kernsymptomen, zolang je maar een minimaal aantal symptomen rapporteert. Uit de resultaten van hoofdstuk 5 bleek echter dat deze benadering niet houdbaar is omdat symptomen juist verschillend samenhangen met depressie. Sommige symptomen bleken, naast de kernsymptomen, erg belangrijk voor depressie en andere symptomen speelden juist een kleinere rol.

In hoofdstuk 6 benadrukten we dat factoren secundair aan depressie er direct voor kunnen zorgen dat symptomen sterk met elkaar samenhangen, wat de bevindingen van data-gedreven studies sterk kan beïnvloeden. Als voorbeeld lieten we zien dat wanneer je zowel af- en toename van gewicht alsook eetlust afzonderlijk analyseert, je automatisch een groep vindt die afname van gewicht en eetlust rapporteert en een andere groep die juist een toename rapporteert. Deze patronen lijken te suggereren dat er verschillende subgroepen van patiënten zijn, maar zijn in werkelijkheid vooral het gevolg van de analysemethode en de manier waarop de symptomen zijn meegenomen. Het is daarom van belang om de juiste statistische modellen te gebruiken om er voor te zorgen dat de kans dat je patronen in de data vindt die kunstmatig zijn te minimaliseren.

In hoofdstuk 7 en 8 gebruikten we bovenstaande kennis om met geavanceerde statistische modellen data-gedreven groepen te vinden van mensen die vergelijkbare patronen van symptomen ervaren. We gebruikten data uit twee grote populatie onderzoeken: Lifelines (73,403 mensen) en NEMESIS (5,583 mensen). We namen hierbij niet aan dat bestaande diagnostische systemen kloppen. In de modellen bekeken we zowel depressie als angst symptomen, namen we extra variabelen mee die klinisch relevante informatie gaven over verschillen in sociaal en lichamelijk functioneren en beperkingen op het werk. Daarnaast maakten we geen onderscheid tussen mensen met of zonder psychiatrische diagnose.

De resultaten lieten zien dat er weinig onderscheid was tussen depressie en angst, waarbij we alleen groepen vonden die diagnose-overstijgend zijn met gemengde depressie en angst klachten. Daarnaast bleek het meenemen van sociaal en lichamelijk functioneren en beperkingen in het dagelijks leven erg informatief voor het begrijpen van verschillen tussen depressieve mensen. Tevens vonden we een grote groep die niet voldoet aan huidige diagnostische criteria maar wel degelijk symptomen ervaart en daar last van heeft in het dagelijks leven. De uitkomsten suggereren dat het beter is om een breder classificatiesysteem te gebruiken dat patiënten minder in kleine hokjes indeelt dan het huidige systeem. Daarbij lieten de resultaten zien dat het informatief is om meer klinisch relevante informatie te betrekken dan alleen symptomen in zowel onderzoek als diagnostiek.

HET METEN VAN DEPRESSIE

Depressie is niet direct te meten maar volgt uit de symptomen die iemand ervaart. Om de ernst van depressie vast te stellen worden daarom vragenlijsten gebruikt. Deze vragenlijsten worden gebruikt om de ernst en het beloop van depressie te meten in zowel klinische zorg als in onderzoek. In het onderzoek naar depressie is daarom een belangrijke vraag hoe we depressie (en de ernst ervan) het beste kunnen meten. Wanneer we de verschillen tussen patiënten beter willen begrijpen en nieuwe verbanden met biologische,

genetische en psychologische variabelen willen leggen, hebben we een sterke en valide maat nodig voor depressie. De data-gedreven benadering die in dit proefschrift wordt gehanteerd vertelt ons meer over hoe goed we depressie op dit moment kunnen meten en biedt nieuwe uitgangspunten om verder te zoeken naar mogelijkheden om depressie preciezer te meten.

De vele factoren die ervoor kunnen zorgen dat iemand depressieve symptomen ervaart zorgen ervoor dat de gerapporteerde symptomen niet altijd een goede weerspiegeling zijn van de ernst van depressie. In zowel hoofdstuk 4 en 5 keken we daarom of we depressie beter en preciezer kunnen meten wanneer we de scores op een vragenlijst aanpassen aan de verschillen tussen depressieve symptomen en patiënten. In hoofdstuk 4 corrigeerden we voor verschillen die er zijn tussen groepen die niet indicatief zijn voor depressie, zoals het rapporteren van somatische klachten bij chronisch zieken. In hoofdstuk 5 keken we naar de verschillen tussen symptomen in hoe ze gemeten worden met vragenlijst-items en de daarbij behorende antwoordcategorieën. Hieruit bleek dat de sterkte waarmee symptomen samenhangen met depressie verschillen en sommige symptomen informatiever zijn dan anderen. Optimale scores werden berekend door de symptomen verschillend te wegen aan de hand van deze resultaten.

Het gebruik van optimale scores bleek weinig uit te maken voor het classificeren van depressie. De ruwe totaalscores die altijd gebruikt worden zijn robuust en veel individuele verschillen en secundaire factoren hebben maar een kleine impact op het totaalplaatje, in het bijzonder wanneer data in grote steekproeven worden geanalyseerd. Mensen die ernstig depressief zijn zullen vrijwel altijd hoog scoren en mensen die niet depressief zijn scoren vrijwel altijd laag. Het gebruik van alternatieve wegingen verandert hier weinig aan. In onderzoek kunnen kleine verschillen echter wel relevant zijn. De effecten van individuele biologische en genetische factoren op psychiatrische stoornissen zijn klein. Hierdoor kan een verbeteringen in onze meting van depressie, hoe marginaal ook, erg relevant zijn en ons beter in staat stellen om kleine, maar relevante verbanden aan te tonen.

Daarnaast suggereren de resultaten ook dat er ruimte is voor verbetering in het meten van depressie met behulp van vragenlijsten. Uit hoofdstuk 3 blijkt bijvoorbeeld dat het misschien beter is om somatische klachten niet uit te vragen bij hartpatiënten wanneer je depressie wilt meten. In hoofdstuk 5 vonden we dat niet elk symptoom even goed gemeten wordt met meerdere ernst categorieën en voor een aantal symptomen alleen de aanwezigheid van het symptoom (ja/nee) van belang lijkt.

KLINISCHE TOEPASSING

Veel mensen met depressieve klachten gaan eerst naar de huisarts. Uit recente cijfers blijkt dat 52 procent van de mensen met psychische klachten op een wachtlijst terecht komen bij de huisarts. Deze tijd kan al gebruikt worden om informatie te verzamelen over de klachten van de patiënt. Data-gedreven methodes kunnen helpen met het samenvatten van deze informatie, die gebruikt kan worden om ervoor te zorgen dat de juiste zorg bij de juiste patiënt terecht komt. De resultaten van hoofdstuk 7 suggereren dat er een grote groep mensen is die stemmings- en angstklachten hebben en daar beperkingen van ondervinden, maar op een niveau waarbij ze niet voldoen aan criteria voor een klinische diagnose. Een in de praktijk toegepast data-gedreven systeem zou een indicatie kunnen geven voor de juiste zorg, waarbij mensen met milde klachten naar laagdrempelige zelfhulp en internettherapie worden verwezen, en mensen met ernstigere klachten naar specialistische zorg. Een dergelijk systeem kan hulpverleners helpen bij het bieden van zorg op maat, ook aan mensen die niet voldoen aan diagnostische criteria maar wel last hebben van hun klachten en baat hebben bij laagdrempelige zorg.

De empirische benadering van dit proefschrift eindigt in hoofdstuk 9, waar we keken naar de toegevoegde waarde van de data-gedreven statistische tool uit hoofdstuk 2 en 3 in een klinische setting. Wij voerden een pilot studie uit waar we keken of het identificeren van patiënten met inconsistente antwoordpatronen op een depressie vragenlijst klinische waarde heeft, en of het terugkoppelen van deze informatie aan hulpverleners bruikbaar kan zijn voor hen. De bevindingen lieten zien dat van de twintig inconsistent rapporterende patiënten in de studie, bij negentien een duidelijke oorzaak was aan te wijzen, die veelal voortkwam uit complexiteit van de psychische klachten die mensen hebben in de specialistische zorg. Verder liet de studie zien dat behandelaren deze extra informatie klinisch nuttig vonden. Een melding zou kunnen aantonen of bevestigen dat de ingevulde vragenlijst geen goede schatting van de ernst van depressie geeft. Ook zou het bruikbaar kunnen zijn om de aandacht van een behandelaar te trekken en hem aan te sporen om de gerapporteerde symptomen nader te bekijken en zou dat aanleiding kunnen geven om hier in een gesprek met de patiënt over te beginnen.

Behandelaren maken klinische beslissingen op basis van de informatie die ze verzamelen gedurende de behandeling van een patiënt. In dit proces kunnen data-gedreven tools een extra bron van informatie zijn voor behandelaren bovenop de informatie die ze nu gebruiken. Data-gedreven systemen moeten dan ook niet gezien worden als vervangende systemen, maar juist als ondersteunende systemen waarbij informatie uit de beschikbare data wordt samengevat op een bruikbare manier en wordt teruggekoppeld aan de behandelaar.

ALGEHELE CONCLUSIE

Dit proefschrift laat zien dat data-gedreven statische methodes zeer bruikbaar zijn in het onderzoek naar depressie, en dat het inzetten van data-gedreven tools de beschikbare klinische informatie over een patiënt kan vergroten, wat uiteindelijk de zorg kan verbeteren. De bevindingen in dit proefschrift gaven weinig empirische evidentie voor de diagnose depressie zoals hij nu wordt gehanteerd. Onderzoek naar depressie kan meer resultaat op leveren wanneer er breder wordt gekeken dan depressieve symptomen, en de diagnose depressie wordt los gelaten.

DANKWOORD

Normaal gesproken eindigt een dankwoord met een aantal zinnen over partner of gezin. Aangezien het volgende beter strookt met het begin van mijn promotietijd, wil ik allereerst jou bedanken Eva. Mijn motivatie komt van jou. Wat onschuldig begon bij een paar eendjes werd al snel iets serieuzer. Ik kan me geen mooier leven voorstellen dan dat met jou… jou en Lea, lief meisje.

Pap en Mam, bedankt voor het voortdurend stimuleren en het helpen groeien van een onverzadigbare honger naar antwoorden, in een verder zorgeloos bestaan. Lieve Mike, lieve Anique, bedankt dat jullie er altijd voor mij zijn.

Ik ben erg dankbaar voor de goede en leuke begeleiding tijdens dit promotietraject door Professor Peter de Jonge, Professor Rob Meijer, en Klaas Wardenaar. Beste Peter, jij bent de voornaamste reden waarom dit promotietraject direct vanaf het begin aan leuk was. Bedankt voor de vrijheid en creativiteit die ervoor zorgde dat ik het altijd interessant en leerzaam vond. Beste Rob, zonder jou was ik geen psychometrie gaan doen en zonder jou ook geen promotietraject, dank daarvoor. Ik heb altijd veel plezier in onze gesprekken en vond het erg leuk dat je langs kwam in Los Angeles. Ik hoop nog lang met je te kunnen samenwerken. Beste Klaas, van jou heb ik wellicht nog het meeste geleerd. Bedankt voor de leuke begeleiding en de vele discussies. Hopelijk volgen er meer.

Ik wil Professor Jeroen Vermunt bedanken voor het warme welkom dat ik kreeg bij mijn bezoek aan hem en zijn groep aan de Tilburg University. Jeroen, ik heb nog nooit zoveel geleerd in korte tijd. Bedankt voor het delen van je kennis en inzichten.

I would like to thank Professor Reise and his group at the University of California, Los Angeles for having me. Steve, thank you for sharing your wise and insightful ideas. To share an office with you was truly an honor. To be locked down with you in that same office during a campus shooting was an experience I will never forget.

Al mijn collega's en coauteurs bedankt. Jullie zorgen voor een ambitieuze en inspiratievolle afdeling waarin het makkelijk is om enthousiast te zijn voor onderzoek. Iedereen van de VICI groep bedankt. Bart, Rei, Stijn en Mara het was leuk om tegelijk als PhD studenten te werken aan vergelijkbare vragen. Daarnaast wil ik iedereen van HoeGekIsNL bedanken, een bijzonder leuk en ambitieus project waaraan ik met veel plezier heb bijgedragen tijdens mijn promotietijd, bedankt Bertus, Lian, Frank, Ando, Maria, Stijn, Evelien, Marieke, Hanneke, Elske, Klaas en Peter.

CURRICULUM VITAE

Rob Wanders was born on 21 June 1986 in Veendam, the Netherlands. After secondary education he started his study Econometrics in 2005 (propaedeutic diploma) and Psychology in 2006 (bachelor degree) at the University of Groningen. During his Bachelor studies he came in contact with Psychometrics, and started a research master in Psychometrics and Quantitative Psychology. During his master he worked as a research assistant for Professor Rob Meijer.

After he had received his master's degree in Psychology, he started his PhD research in 2012 at the Interdisciplinary Center Psychopathology and Emotion regulation (ICPE), under supervision of Professor Peter de Jonge, Professor Rob Meijer, and dr. Klaas Wardenaar on the VICI project 'Deconstructing Depression' where he investigated an empirical approach on depression, which resulted in this thesis. During his PhD project he presented his research at several national and international conferences. He followed several courses among which a masterclass from Professor Von Davier at the University of Oxford. In 2014, he worked with Professor Jeroen Vermunt for a brief period at the Department of Methodology and Statistics, University of Tilburg. In 2016, he stayed for two months at the University of California, Los Angeles to work with Professor Steve Reise. In July 2016, Rob started working as a post-doctoral researcher on a large European project about Comorbid Conditions of ADHD under the supervision of under the supervison of dr. Catharina Hartman, and part time for Professor Peter de Jonge on developmental psychopathology.

LIST OF PUBLICATIONS

Wigman JTW, Wardenaar KJ, **Wanders RBK**, Booij SH, Jeronimus BF, van der Krieke L, Wichers MC, de Jonge, P. (in press). Dimensional and discrete variations on the psychosis continuum in a Dutch crowd-sourcing population sample. *European Psychiatry*.

Wanders RBK, Van Loo HM, Vermunt JK, Meijer RR, Hartman CA, Schoevers RA, Wardenaar KJ, de Jonge P (2016). Casting wider nets for anxiety and depression: disability-driven cross-diagnostic subtypes in a large cohort. *Psychological Medicine*, 46(16), 3371-3382.

van Loo HM, **Wanders RBK**, Wardenaar KJ, Fried EI (2016). Problems with latent class analysis to detect data-driven subtypes of depression. *Molecular Psychiatry*.

Wardenaar KJ, **Wanders RBK**, de Jonge P (2016). Meetinstrumenten voor depressie. In A.H. Schene, B. Sabbe, H.G. Ruhé, P. Spinhoven (Eds). Handboek Depressieve Stoornissen. Uitgeverij De Tijdstroom.

Bos, EH, **Wanders RBK** (2016). Group-level symptom networks are catchy but deceptive. *JAMA Psychiatry,* 73, 411-411.

van der Krieke L, Jeronimus BF, Blaauw F, **Wanders RBK**, Emerencia AC, Schenk H, ... de Jonge P (2015). HowNutsAreTheDutch (HoeGekIsNL): A crowdsourcing study of mental symptoms and strengths. *International journal of methods in psychiatric research*, 25, 123-144.

Wanders RBK, Wardenaar KJ, Penninx BWJH, Meijer RR, de Jonge P (2015). Data-driven atypical profiles of depressive symptoms: Identification and validation in a large cohort. *Journal of Affective Disorders*, 180, 36-43.

Wardenaar, KJ, **Wanders RBK**, Roest AM, Meijer RR, de Jonge P (2015). What does the beck depression inventory measure in myocardial infarction patients? a psychometric approach using item response theory and person-fit. *International Journal of Methods in Psychiatric Research*, 24, 130-142.

Wanders RBK, Wardenaar KJ, Kessler RC, Penninx BWJH, Meijer RR, de Jonge P (2015). Differential reporting of depressive symptoms across distinct clinical subpopulations: What DIFference does it make?. Journal of Psychosomatic Research, 78(2), 130-136.

Meijer RR, Tendeiro JN, **Wanders RBK** (2014). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. Revicki (Eds.). Handbook of item response theory modeling: Applications to typical performance assessment. (pp. 85-110). Routledge.

SUBMITTED FOR PUBLICATION

Wanders RBK, Reise SP, Wardenaar KJ, Haviland MG, de Jonge P, Meijer RR. Investigating the equal weighting and validity of response categories in measuring individual depressive symptoms using the nominal response model, *submitted.*

Wardenaar KJ, **Wanders RBK**, ten Have M, de Graaf R, de Jonge P. Patterns and dimensionality of depressive and anxiety symptomatology in the general population, *under review.*

Wanders RBK, Meijer RR, Ruhé HG, Sytema S, Wardenaar KJ, de Jonge P. Person-fit Feedback on Inconsistent Symptom Reports in Clinical Depression Care, *under review.*